SIGNAL IS HARDER TO LEARN THAN BIAS: DEBIASING WITH FOCAL LOSS

Moritz Vandenhirtz, Laura Manduchi, Ričards Marcinkevičs and Julia E. Vogt
Department of Computer Science, ETH Zurich
{moritz.vandenhirtz, laura.manduchi, ricardsm, julia.vogt}@inf.ethz.ch

ABSTRACT

Spurious correlations are everywhere. While humans often do not perceive them, neural networks are notorious for learning unwanted associations, also known as biases, instead of the underlying decision rule. As a result, practitioners are often unaware of the biased decision-making of their classifiers. Such a biased model based on spurious correlations might not generalize to unobserved data, leading to unintended, adverse consequences. We propose Signal is Harder (SiH), a variational-autoencoder-based method that simultaneously trains a biased and unbiased classifier using a novel, disentangling reweighting scheme inspired by the focal loss. Using the unbiased classifier, SiH matches or improves upon the performance of state-of-the-art debiasing methods. To improve the interpretability of our technique, we propose a perturbation scheme in the latent space for visualizing the bias that helps practitioners become aware of the sources of spurious correlations.

1 INTRODUCTION

The generalization capability of deep neural networks (DNN) highly depends on the quality of the training data. If spurious correlations are present, the model might ignore the intrinsic *signal attributes* while still performing reasonably well in classification tasks. However, such a biased model will not be robust and will not generalize outside the training distribution. To increase the trustworthiness of machine learning algorithms and prevent unwanted consequences, it is crucial to avoid deploying biased models (Geirhos et al., 2020). Thus, there has been an increased interest in the community to mitigate this problem. While many methods assume and utilize an observed variable that captures the source of bias for each data point, recently, some effort has been made to alleviate this prohibitive assumption (Nam et al., 2020).

For example, consider a dataset comprising of images of vehicles. A DNN might implicitly use the *bias attribute* "sky" as a shortcut for classifying planes because most images of airplanes are shot while they are in the air. Throughout the paper, we will call samples *bias-aligned* when their bias attributes are strongly correlated with the label. Here, leveraging the bias as a decision rule leads to the correct predicted label, e.g. airplane in the sky. Conversely, *bias-conflicting* data points are the samples for which the biased decision rule leads to the wrong prediction, e.g. aircraft in the hangar.

Recent efforts by Nam et al. (2020) aim to eliminate the need for an observed variable that captures the source of bias for each data point. They assume that malignant bias attributes are easier to learn than the underlying signal. Based on this easy-to-learn assumption, they train a biased classifier that focuses on the easy, bias-aligned samples. Simultaneously, they train an unbiased classifier by upweighting the remaining hard, bias-conflicting samples. We propose an alternative reweighting for the unbiased classifier based on the focal loss (Lin et al., 2017) that does not require the previously utilized subtle, distorting stability measures. We motivate the usage of this loss function through the easy-to-learn assumption, which infers that signal is harder to learn than bias.

In addition, we extend the literature by integrating a variational autoencoder (VAE) (Kingma & Welling, 2014) into the model. At inference time, this allows us to make use of latent perturbations to remove the biasing attributes from the embeddings, which we then feed to the decoder to visualize debiased images. Comparing these images with the original reconstructions can help practitioners uncover unknown biases.

Contribution We propose a novel reweighting scheme, coined Signal is Harder (SiH), for training an unbiased classifier.[1] Due to the lack of labels for the unknown bias, SiH exploits the assumption that signal is harder to learn than bias and utilizes a reweighting based on the well-established focal loss (Lin et al., 2017). We show that this direct mechanism improves the debiasing capabilities compared to the existing, more complex reweighting scheme by Nam et al. (2020). Additionally, by training a VAE simultaneously with the classifiers, the unknown bias can be visualized in the reconstructions. For this, the proposed algorithm perturbs the latent bias embeddings at inference time to remove the bias without creating artifacts in the reconstructions. We improve upon previous methods as our minimal perturbation does not change other aspects of the reconstruction, unambiguously unveiling the unknown spurious attribute.

2 Related work

Separating samples by difficulty Recent works separate bias-conflicting from bias-aligned samples to train an unbiased classifier (Nam et al., 2020; Lee et al., 2021; Kim et al., 2021). This separation can be achieved by differentiating data points through the difficulty of predicting their label. In a standard classification setting, Zhang & Sabuncu (2018) propose the Generalized Cross Entropy (GCE) loss to reduce the weight on samples whose labels are hard to predict:

$$GCE(\hat{y}, y) = \frac{1 - \hat{y}^q}{q}, \tag{1}$$

where $\hat{y}$ is the predicted probability of the correct label y according to the classifier, and $q \in (0, 1]$ is a hyperparameter to control the strength of emphasis. The GCE is best understood by inspecting its derivative $\frac{\partial GCE(\hat{y},y)}{\partial \boldsymbol{\theta}} = \hat{y}^q \frac{\partial CE(\hat{y},y)}{\partial \boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ are the learnable neural network parameters. This loss upweighs samples that the classifier already predicts well, ignoring samples for which the current decision rule does not work. Contrary to the GCE loss, the Focal Loss (FL) by Lin et al. (2017) puts more focus on hard, misclassified examples:

$$FL(\hat{y}, y) = (1 - \hat{y})^q CE(\hat{y}, y) \tag{2}$$

With this reweighting scheme, the samples whose labels are hard to predict are upweighted such that the classifier does not ignore the samples for which finding a decision rule is a hard problem.

Debiasing without supervision Previous works focused on predictions with respect to known sensitive attributes (Sagawa et al., 2020; Edwards & Storkey, 2016; Kim et al., 2019), which are often difficult to retrieve. For this reason Nam et al. (2020) propose LfF, a new approach to debias a classifier, which does not require bias attributes. They assume that bias is only malignant if it is easier to learn than the true signal attribute and leverage the GCE loss to focus on the easy, bias-aligned samples to train a biased classifier. Simultaneously, they train an unbiased classifier, designed to learn the true, underlying signal. For this they upweigh the bias-conflicting samples, i.e. the data points for which the bias can not be utilized to predict the label, by the relative difficulty score (RDS)

$$RDS(\hat{y}_s, \hat{y}_b, y) = \frac{CE(\hat{y}_b, y)}{CE(\hat{y}_s, y) + CE(\hat{y}_b, y)}, \tag{3}$$

where $\hat{y}_s$ and $\hat{y}_b$ are the predicted probabilities of the correct label y according to the unbiased and biased classifier, respectively. However, before inserting the CE terms into the above formula, they apply an empirically motivated exponential moving average and a class-wise normalization by the maximum CE to each term. In the following section, we will propose an enhanced upweighting mechanism for the unbiased classifier that does not require weight-distorting stability measures.

Lee et al. (2021) extend the method of Nam et al. (2020) by additionally swapping the learned latent bias embeddings of different inputs to decouple the bias from the label. At inference time, they train a decoder to visualize the embeddings with and without swapped bias, such that the unknown bias can be discovered by analyzing the differences between the two reconstructions. To avoid misleading artifacts in the visualization, we will propose a more conservative perturbation, which relies on a VAE trained simultaneously with the classifiers. Further discussion can be found in Appendix A.

[1] Our code is publicly available at `https://github.com/mvandenhi/Signal-is-Harder`

3 METHOD

We propose a debiasing algorithm, coined Signal is Harder (SiH), consisting of a VAE-based architecture and a new weighting mechanism for training the unbiased classifier. The VAE uses two encoders to map the input into signal and bias embeddings, which are concatenated and passed through the decoder to reconstruct the original input. Additionally, we train an unbiased and biased classifier on signal and bias embeddings, respectively. The biased classifier is trained by upweighting bias-aligned samples through the GCE loss. In contrast, the unbiased classifier is trained by upweighting bias-conflicting samples through our novel focal-loss-based weighting scheme, which we will introduce in the next paragraph. The generative nature of the model allows us to produce bias visualizations that help discover the unknown source of bias. We depict the proposed model structure in Figure 1.

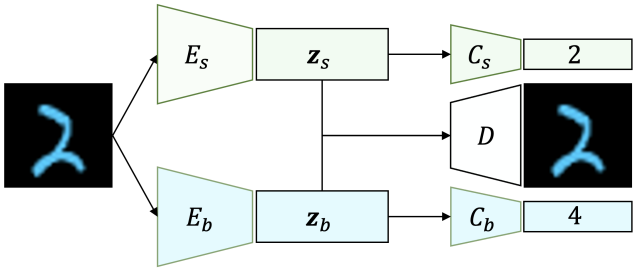

Figure 1: Graphical overview of our model's structure for a bias-conflicting image. The input $\mathbf{x}$ is passed through the signal and bias encoders E_s and E_b to obtain the latent signal and bias embeddings $\mathbf{z}_s$ and $\mathbf{z}_b$, which in this example should be the digit *two* and the color *blue*, respectively. These representations are then passed through their respective classifier C_s and C_b to predict the label. Lastly, $\mathbf{z}_s$ and $\mathbf{z}_b$ are concatenated and passed through the decoder D to reconstruct the image.

Reweighting by focal loss We hereby introduce a new reweighting scheme, aiming to utilize the easy-to-learn assumption not only for the biased but also for the unbiased classifier. Similar to the GCE that upweighs bias-aligned samples for the biased classifier, we want to utilize a mirrored loss function that upweighs the remaining bias-conflicting samples for the unbiased classifier.

The aforementioned reasons motivate the inclusion of the focal loss for training the unbiased classifier. We use this loss to identify samples for which the biased classifier struggles to predict the correct class and emphasize these presumably bias-conflicting samples by upweighting them when training the unbiased classifier. As the information learned by the biased classifier should leverage the unbiased classifier, but not the other way around, we detach the weighting factor from the computational graph during backpropagation and obtain the following update for the unbiased classifier:

$$\frac{\partial \mathcal{L}_s(\hat{y}_s, \hat{y}_b, y)}{\partial \boldsymbol{\theta}_s} = (1 - \hat{y}_b)^q \frac{\partial CE(\hat{y}_s, y)}{\partial \boldsymbol{\theta}_s}, \tag{4}$$

where $q \in (0, 1]$ is a hyperparameter controlling the strength of emphasis. With this loss, we exploit that bias-conflicting samples are hard to learn for a biased classifier. By focusing on these data points, the unbiased classifier is forced to learn the signal, as here, leveraging the bias does not lead to the correct prediction. Most importantly, the straightforward integration of the focal loss for training the unbiased classifier removes the need for weight-distorting stability measures.

Latent adversarial perturbation To make practitioners aware of the unknown spurious correlations in a dataset, we propose a visualization approach by perturbing the bias at inference time. We perturb the bias embeddings, such that the bias contained within is removed, and reconstruct the debiased image. We argue that such a perturbation should be as small as possible, such that no artifacts are created in the process since a practitioner needs to consider every change to the input as a potential bias. To achieve this, we adapt and utilize the adversarial perturbations from Deepfool (Moosavi-Dezfooli et al., 2016). This algorithm is designed to find a minimal perturbation to the input that fools the classifier into predicting the wrong class. Thus, we perturb the bias representations such that the biased classifier can no longer predict the correct class; effectively, the perturbation removes the bias from the bias embeddings.

Table 1: Unbiased test set accuracy + standard deviation in %. The method with the significantly highest accuracy is denoted in **bold**. Otherwise, insignificantly different methods are underlined.

Dataset	Ratio	Vanilla	LfF	DisEnt	SiH
Colored MNIST	20%	**94.92** ± 0.24	70.18 ± 4.19	90.94 ± 1.46	85.24 ± 1.60
	10%	**91.24** ± 0.26	81.99 ± 5.01	89.12 ± 1.44	85.35 ± 1.23
	5%	85.48 ± 0.50	81.18 ± 2.94	85.54 ± 2.49	86.14 ± 1.78
	2%	73.28 ± 0.56	76.97 ± 2.49	82.38 ± 1.68	83.80 ± 1.28
	1%	59.41 ± 0.39	68.91 ± 5.01	76.33 ± 3.41	**80.03** ± 2.04
	0.5%	43.70 ± 0.83	60.42 ± 2.72	63.98 ± 4.78	**71.63** ± 2.49
Corrupted CIFAR-10	20%	67.57 ± 0.41	64.50 ± 2.17	60.99 ± 5.84	66.75 ± 1.34
	10%	57.11 ± 0.76	59.29 ± 3.16	53.47 ± 4.43	61.26 ± 2.06
	5%	46.89 ± 0.78	55.77 ± 2.33	46.40 ± 5.81	55.63 ± 1.54
	2%	34.90 ± 0.81	47.26 ± 1.56	36.98 ± 4.43	43.66 ± 1.81
	1%	28.22 ± 0.73	**39.39** ± 2.16	31.22 ± 2.69	35.17 ± 1.19
	0.5%	22.26 ± 1.03	30.04 ± 1.67	31.97 ± 3.34	27.30 ± 2.04

Having trained a VAE, we can use the decoder at inference time to visualize the perturbed bias embeddings together with the unchanged signal representation. By comparing the original reconstruction with the debiased visualization, it is possible to identify the spurious correlations in the image. In contrast to DisEnt, we train the decoder simultaneously with the classifiers to encode all image-relevant information in the latent representations, thus, supporting the unbiased classifier in finding the signal attributes and improving reconstruction quality.

4 EXPERIMENTS

To compare the performance of our method, SiH, with previous works, we evaluate it on Colored MNIST (Kim et al., 2019) and Corrupted CIFAR-10 (Hendrycks & Dietterich, 2019) with a varying percentage of bias-conflicting images during the training. For a detailed description of the datasets, we refer to Appendix C. We determine three baselines to which we compare the proposed approach. The first baseline we implement is a Vanilla model consisting of one encoder and classifier, which measures the standard performance without any debiasing scheme. The second model we compare the proposed approach to, is LfF from Nam et al. (2020). Lastly, we compare SiH to DisEnt by Lee et al. (2021), a recently proposed state-of-the-art debiasing algorithm, which also visualizes the bias.

4.1 QUANTITATIVE EVALUATION

Comparison on test sets In Table 1, we show the performance of all models on the unbiased test set of Colored MNIST and Corrupted CIFAR-10. The estimates differ from the values presented in the baseline papers (Nam et al., 2020; Lee et al., 2021) because we also vary random seeds over dataset generation instead of only over the weight initialization.

We observe that Vanilla outperforms the debiasing algorithms for the 10% and 20% cases of Colored MNIST. Thus, the easy-to-learn assumption is likely not fulfilled for these training sets. The debiasing methods show their benefit only for a lower amount of bias-conflicting samples. Here, SiH outperforms or at least matches all baselines, while DisEnt is the runner-up. Especially for the 0.5% setting, there is a considerable gap in performance between our and other methods.

For Corrupted CIFAR-10, the best-performing models are LfF and SiH. We observe that for higher percentages of bias-conflicting samples, SiH is better than LfF, while for lower proportions, the opposite is the case. DisEnt seems to be generally worse than the other debiasing methods. Finally, Vanilla performs worse than the debiasing methods except for the 20% case. Thus, the debiasing methods present an improvement over a standard empirical risk minimizer.

Focal loss vs. RDS weighting In Table 2, we show the performance on the unbiased test set of Colored MNIST and Corrupted CIFAR-10 using two different ways of weighting the samples for training the unbiased classifier. While SiH stands for our proposed method, in SiH$_{RDS}$, we utilize the RDS proposed by Nam et al. (2020) when reweighting the data points. The results suggest that on average our proposed reweighting mechanism significantly increases performance while the inclusion of the VAE leads to an accuracy-interpretability tradeoff. Additionally, the focal loss reduces the variability in accuracy across multiple runs.

Figure 2: A random collection of bias visualizations for Colored MNIST. The randomly selected images are varied over random seeds and the percentage of bias-conflicting images in the training set.

Table 2: Unbiased accuracy + standard deviation in % for Colored MNIST and Corrupted CIFAR-10.

Dataset	Ratio	SiH$_{RDS}$	SiH
Colored MNIST	20%	80.15 ± 5.27	**85.24** ± 1.60
	10%	86.13 ± 3.07	85.35 ± 1.23
	5%	84.10 ± 3.03	86.14 ± 1.78
	2%	79.38 ± 2.37	**83.80** ± 1.28
	1%	74.22 ± 3.21	**80.03** ± 2.04
	0.5%	64.17 ± 5.74	**71.63** ± 2.49
Corrupted CIFAR-10	20%	64.09 ± 5.64	66.75 ± 1.34
	10%	56.89 ± 5.09	**61.26** ± 2.06
	5%	51.25 ± 4.10	**55.63** ± 1.54
	2%	38.22 ± 3.77	**43.66** ± 1.81
	1%	31.64 ± 2.50	**35.17** ± 1.19
	0.5%	24.59 ± 1.84	**27.30** ± 2.04

Overall, the quantitative results show that the proposed reweighting scheme improves performance. For settings where the easy-to-learn assumption is likely to be fulfilled, SiH shows promising results compared to baselines. The ablation study shows that integrating our reweighting for the unbiased classifier is critical in improving its accuracy.

4.2 Qualitative evaluation

Figure 2 displays the bias visualization from DisEnt and SiH for a few randomly selected images. Additionally, in Figure 7 of Appendix D, we show the random bias visualizations for Corrupted CIFAR-10. We will not analyze the latter images further, as here, signal and bias are not disentangled well enough for visualizing the bias for either method.

For Colored MNIST, the swapping of DisEnt perturbs the bias representations so strongly that this also leads to an unwanted change in the digit. This change is likely due to the bias and signal representations not being perfectly disentangled. Thus, the leftover signal in the bias dimensions gets swapped too. On the other hand, SiH does not perturb the digit while regularly perturbing the color. However, due to the weaker magnitude of change, our approach sometimes does not visibly change the image.

SiH is more conservative when generating perturbations, which is advantageous for visualizing bias in realistic cases where learned signal and bias embeddings are not perfectly disentangled. Although our changes are more subtle, we believe that, for a practitioner, our perturbation method should be preferred, as it does not induce artifacts, which otherwise have to be considered as a possible bias.

5 Conclusion and future work

In the presence of bias, a classifier often leverages these spurious correlations rather than the underlying signal. The application of such an algorithm can have adverse consequences in critical situations. This work advances the research in building unbiased deep learning models by investigating a novel reweighting scheme. We propose SiH, which trains a bias classifier to be as biased as possible and simultaneously trains an unbiased classifier by upweighting samples for which the biased decision rule fails to predict the correct labels. We show that the proposed weighting factor based on the focal loss can match or outperform existing works. Additionally, by training a generative model, users are able to visualize and identify the bias at inference time. For this, the proposed approach leverages latent adversarial perturbations that do not introduce undesirable artifacts.

Future work Although SiH has demonstrated its effectiveness on simple datasets, its efficacy on more challenging datasets and other modalities requires further investigation. For this, it is crucial to use more expressive generative models such as generative adversarial networks (Goodfellow et al., 2020) or diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020). Moreover, while SiH consists of established individual components, their combination is not rigorously derived. In fact, the entire field would profit from greater mathematical rigor, beginning with the establishment of a theoretical definition of what constitutes the "ease of learning".

REFERENCES

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pp. 528–539. PMLR, 2020.

Rémi Cadène, Corentin Dancette, Hédi Ben-Younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/51d92be1c60d1db1d2e5e7a07da55b26-Abstract.html.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4067–4080, 2019.

Luke Darlow, Stanisław Jastrzebski, and Amos Storkey. Latent adversarial debiasing: Mitigating collider bias in deep neural networks. *arXiv preprint arXiv:2011.11486*, 2020.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations*, 2016.

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

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020. doi: 10.1145/3422622. URL https://doi.org/10.1145/3422622.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9012–9020, 2019.

Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14992–15001, 2021.

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

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6114.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, 2009.

Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *Advances in Neural Information Processing Systems*, 2021.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pp. 2999–3007. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.324. URL https://doi.org/10.1109/ICCV.2017.324.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Debiasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*. OpenReview.net, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2256–2265. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11895–11907, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pp. 6306–6315, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*, 2019.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, pp. 8792–8802, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/f2925f97bc13ad2852a7a551802feea0-Abstract.html.

A FURTHER RELATED WORK

Debiasing without explicit supervision To train an unbiased classifier, several existing approaches (Bahng et al., 2020; Clark et al., 2019; Wang et al., 2019; Cadène et al., 2019) leverage some form of implicit knowledge about the bias present in the dataset. They use this insight to design and train a model architecture susceptible to the specific bias attribute. They then train an unbiased model to learn decision rules different from the biased model. For example, Bahng et al. (2020) capture texture bias in image classification by training a convolutional neural network with a small receptive field. Simultaneously, they train an unbiased model by forcing it to learn representations that are independent of the biased ones.

Debiasing without implicit supervision Liu et al. (2021) slightly adapt the idea of Nam et al. (2020). Instead of training both classifiers simultaneously, they divide the training into two stages. First, the biased classifier is trained. Second, they train the unbiased classifier and upweigh all samples that were misclassified by the biased classifier. Hence, they also try to focus on training on the bias-conflicting samples.

DisEnt by Lee et al. (2021) builds upon Nam et al. (2020) by utilizing their training scheme to create a disentangled representation useful for feature augmentation. As motivation, they show that the diversity of training samples is an important factor in training. Instead of training the unbiased classifier only on the sparse bias-conflicting samples, they try to synthesize additional samples for which using the bias as decision rule does not work. Their algorithm works as follows: First, Lee et al. (2021) train the base structure from Nam et al. (2020) using GCE and RDS to create disentangled representations where signal and bias dimensions can be separated. After a predetermined number of updates, they start to swap the bias dimensions of different samples to synthesize representations with the same signal but different bias. Hence, making the bias unusable as decision rule for the label because it originates from a different sample. They then train their classifiers on those representations as well as on the original samples.

A caveat of Nam et al. (2020) is the absence of a mechanism for visualizing the unknown bias. For this, Lee et al. (2021) propose to train a decoder ex-post to reconstruct the images given their latent representations. By repeating their swapping process and reconstructing the images after the swap, they expect to visualize the bias present in the dataset as they can compare reconstructions with the same signal but swapped bias dimensions. This will work if the learned embeddings are perfectly disentangled but might introduce artifact if they are not.

Further research for debiasing and visualizing the bias has been done by Darlow et al. (2020), which leverage a vector quantized variational autoencoder (van den Oord et al., 2017) and add a simple biased classifier on top of the latent representations. After training, they perturb the latent dimensions such that the biased classifier is as unsure as possible about the label. This perturbed representation is then passed through the decoder to generate images without relevant bias. Consequently, a second, unbiased classifier is trained on these images, which should not contain bias that can be used for predicting the label.

B IMPLEMENTATION DETAILS

In line with previous works (Nam et al., 2020; Lee et al., 2021; Kim et al., 2021), we utilize an MLP for Colored CMNIST for all methods. Each encoder consists of three linear layers with a bottleneck of size 100 for signal and bias, respectively. The decoder is again consisting of three linear layers. As activation function, we use the Rectified Linear Unit (ReLU). For the classifiers, we solely use one linear layer for both datasets. This is because for us, the difficulty of a dataset is determined by the complexity of the connection between latent variables and the realization $\mathbf{x}$ thereof. If we knew the latent variables, inferring the label would be simple. Thus, a single linear layer suffices.

We adapt the encoder-decoder structure for Corrupted CIFAR-10 from an MLP to a CNN, where we use a ResNet18 (He et al., 2016) with bottleneck dimension of 512 for the encoders of all methods and a ResNet18-like decoder.

We do not perform any preprocessing for Colored MNIST. For the preprocessing of Corrupted CIFAR-10, we take random crops consisting of at least 50% of the original image and resize them

to the original 32×32 size. Additionally, we allow horizontal flips of the images and standardize the pixel values over the entire dataset. To calculate the reconstruction loss, we transform the standardized pixel values back into $[0, 1]$ so that its size is comparable among all datasets.

To have visualizations that capture the original image well, for SiH, we upweigh the reconstruction loss by the factor 100. Additionally, we rescale the reconstruction and KL term by dividing through $3L$, where L is the number of pixels, to be invariant to image resolution and number of channels while retaining their relative loss magnitude.

For updating the model weights, we use the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.001 and batch size of 256 for both datasets. The hyperparameter q is chosen to be 0.7 by following the GCE coiners Zhang & Sabuncu (2018).

We perform early stopping and reduce the learning rate when plateauing by computing this loss function on a held-out 10% of the training set. For Colored MNIST, we use an early stop patience of 2 versus 20 for Corrupted CIFAR-10. The patience for the learning rate reduction is one-half of the early stop patience and reduces the learning rate by a factor of 10.

We do not perform hyperparameter tuning on specific settings, as there must not be an unbiased or bias labelled validation set in the setting of unknown bias. We would like to encourage future work to do the same.

For creating visualization of the bias, we adapt Deepfool (Moosavi-Dezfooli et al., 2016) for our purposes. Originally, the algorithm was developed for perturbing pixels in an input image, while we use it for perturbing latent dimensions. Thus, while pixel values need to be clamped in $[0, 1]$, we do not require this. For the distance measure of the perturbation, we use the ℓ_2-norm.

For the bias visualization, we use the trained model of SiH as backbone. This is vital in ensuring that discrepancies in the visualizations of DisEnt and SiH can be attributed solely to the differing visualization techniques. We perturb images for which the biased and unbiased classifiers predict the correct class. With this, we aim to find bias-aligned images, which we can then perturb into neutral images. We randomly sample a different target class and apply our perturbation. For DisEnt, we pick a second image, for which the biased classifier predicts the same sampled target class and swap the bias embeddings.

C Dataset visualizations

To compare the performance of our method SiH with previous works, we evaluate it on two datasets. These datasets consist of a training set for which we define varying percentages of bias-conflicting images to analyse the performance. We assess the debiasing potential of all methods on an unbiased test set where signal and bias are independently and uniformly distributed. For SiH, we do not perform hyperparameter tuning on each framework because an unbiased validation set does not exist in the setting of unknown bias.

The first dataset is Colored MNIST by Kim et al. (2019), which consists of the popular handwritten digit database MNIST (Lecun et al., 1998) synthetically infused with a color bias. To each digit, we randomly assign distinct mean colors, which serve as bias attributes. Hence, the signal $\mathbf{z}_s$ is the digit while the easy-to-learn bias $\mathbf{z}_b$ manifests itself as the color. In Figure 3, we display bias-aligned images, for which leveraging the color as decision rule would lead to the correct label. A minority of samples in the training set consists of bias-conflicting samples, showed in Figure 4, for which the biased decision rule leads to the wrong prediction. Learning to recognize the digit instead of the color is the only valid decision rule with which bias-aligned as well as bias-conflicting samples can be correctly classified. The second dataset we apply SiH to is the Corrupted CIFAR-10 dataset (Hendrycks & Dietterich, 2019). It is based on the standard CIFAR-10 dataset (Krizhevsky & Hinton, 2009), injected with synthetically generated corruptions such as fog, brightness, or saturation for each class. These synthetic biasing perturbations are designed to be as realistic as possible. A collection of randomly selected bias-aligned and bias-conflicting images for both datasets can be found in Figure 5 and Figure 6, respectively.

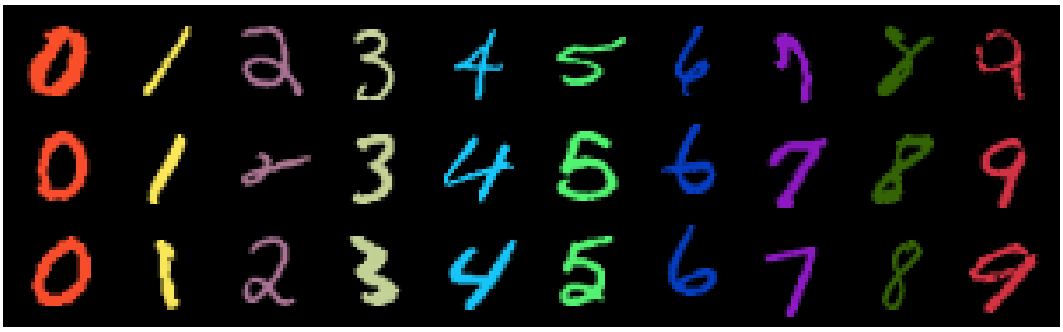

Figure 3: Bias-aligned images of the Colored MNIST dataset. The columns show different digits $\mathbf{z}_s$ with their respective colors $\mathbf{z}_b$ that predominantly manifest in combination.

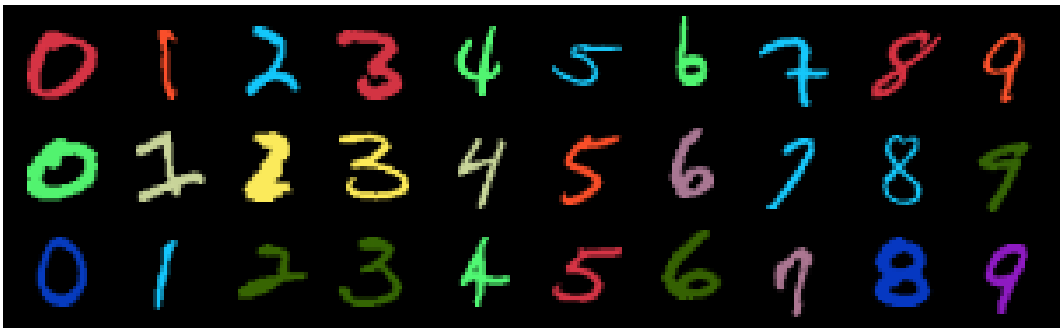

Figure 4: Bias-conflicting images of the Colored MNIST dataset. The columns show different digits $\mathbf{z}_s$ with colors $\mathbf{z}_b$ that are usually not observed together.

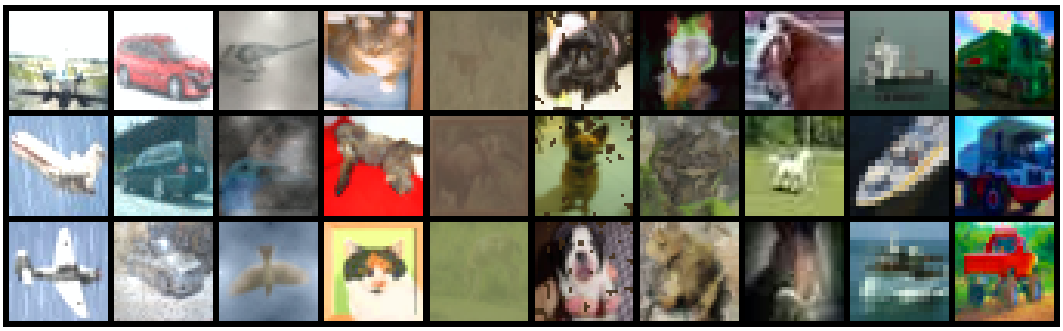

Figure 5: Bias-aligned images of the Corrupted CIFAR-10 dataset. The columns show the different classes $\mathbf{z}_s$ with their respective corruptions $\mathbf{z}_b$ that predominantly manifest in combination. For example, the class birds often has foggy images, while ships are frequently pixelated.

Figure 6: Bias-conflicting images of the Corrupted CIFAR-10 dataset. The columns show the different classes $\mathbf{z}_s$ with corruptions $\mathbf{z}_b$ that are usually not observed together.

D BIAS VISUALIZATIONS FOR CORRUPTED CIFAR-10

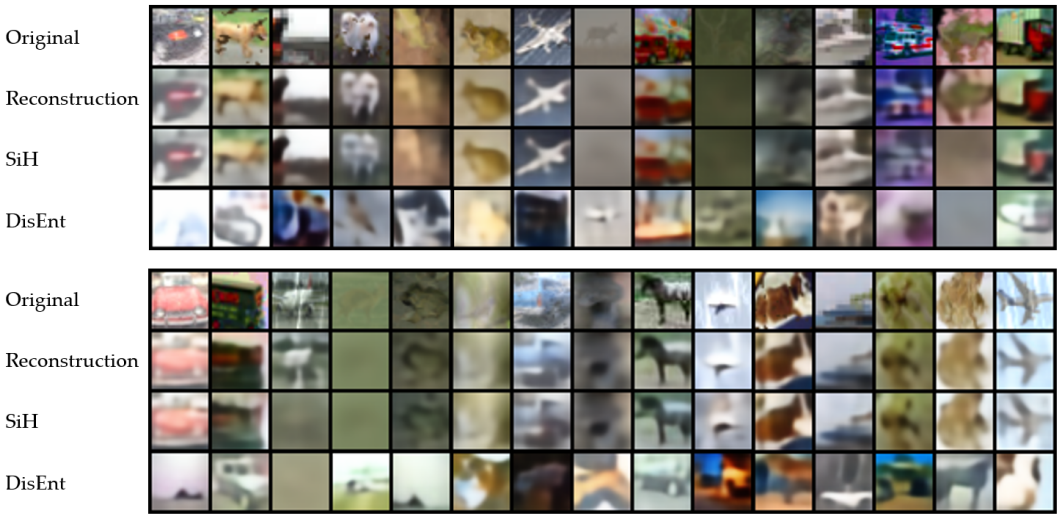

Figure 7: A random collection of bias visualizations for Corrupted CIFAR-10. The randomly selected images are varied over random seeds and the percentage of bias-conflicting images in the training set.