Enhancing Intrinsic Features for Debiasing via Investigating Class-Discerning Common Attributes in Bias-Contrastive Pair

Supplementary Material

This supplementary material offers further analysis of our approach, additional experimental results, the details of the datasets and implementation, limitations, and future work. Sec. A and Sec. B provide the analysis of the bias-negative (BN) score as a loss weight and samples with negative BN score, respectively. Sec. C analyzes the effect of BC samples in $\mathcal{D}^{\text{BN}}$ on debiasing performance. Also, Sec. D compares the recent sample selection methods with ours. Moreover, Sec. E and Sec. F present additional qualitative results regarding the guidance and additional quantitative results, respectively. Sec. G and Sec. H provide the details about the dataset and implementation. Lastly, Sec. I discusses the limitations and future work.

A. Additional analysis of the BN score as a loss weight

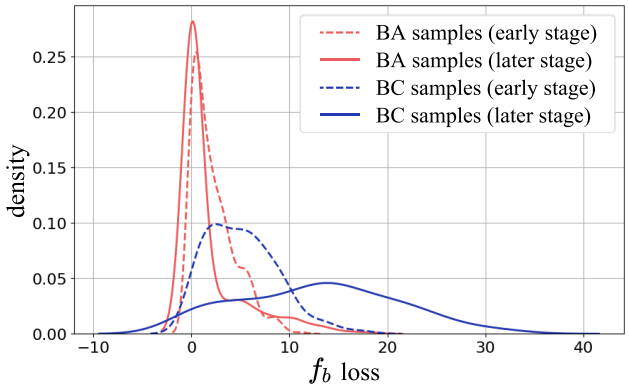


Figure 1. The distributions of f_b's classification loss of samples in $\mathcal{D}^{\text{BN}}_{\text{cand}}$. The red and blue lines denote the losses of BA and BC samples, respectively. The dotted and solid lines indicate the losses at the early and later stages of the training, respectively. Best viewed in color.

As described in Sec. 3.4 in the main paper, we utilize the BN score of $\mathbf{x}^{\text{BN}}$ (*i.e.*, $s(\mathbf{x}^{\text{BN}})$) to reweight the guidance loss $\mathcal{L}_{\text{guide\_sim}}$ and the BN loss $\mathcal{L}_{\text{BN}}$. The BN score as a loss weight is designed to upweight the losses when bias-conflicting (BC) samples are selected as $\mathbf{x}^{\text{BN}}$, which further encourages our IE weight to enhance the intrinsic features. For verification, we present that the BN score has a much larger value on the BC samples compared to bias-aligned (BA) samples during the training in Fig. 2 in the main paper.

Since $s(\mathbf{x}^{\text{BN}})$ has a larger value when the current f_b loss of $\mathbf{x}^{\text{BN}}$ is larger than that of the early stage of training, the results imply that the f_b loss of BC samples largely increases as training proceeds compared to BA samples.

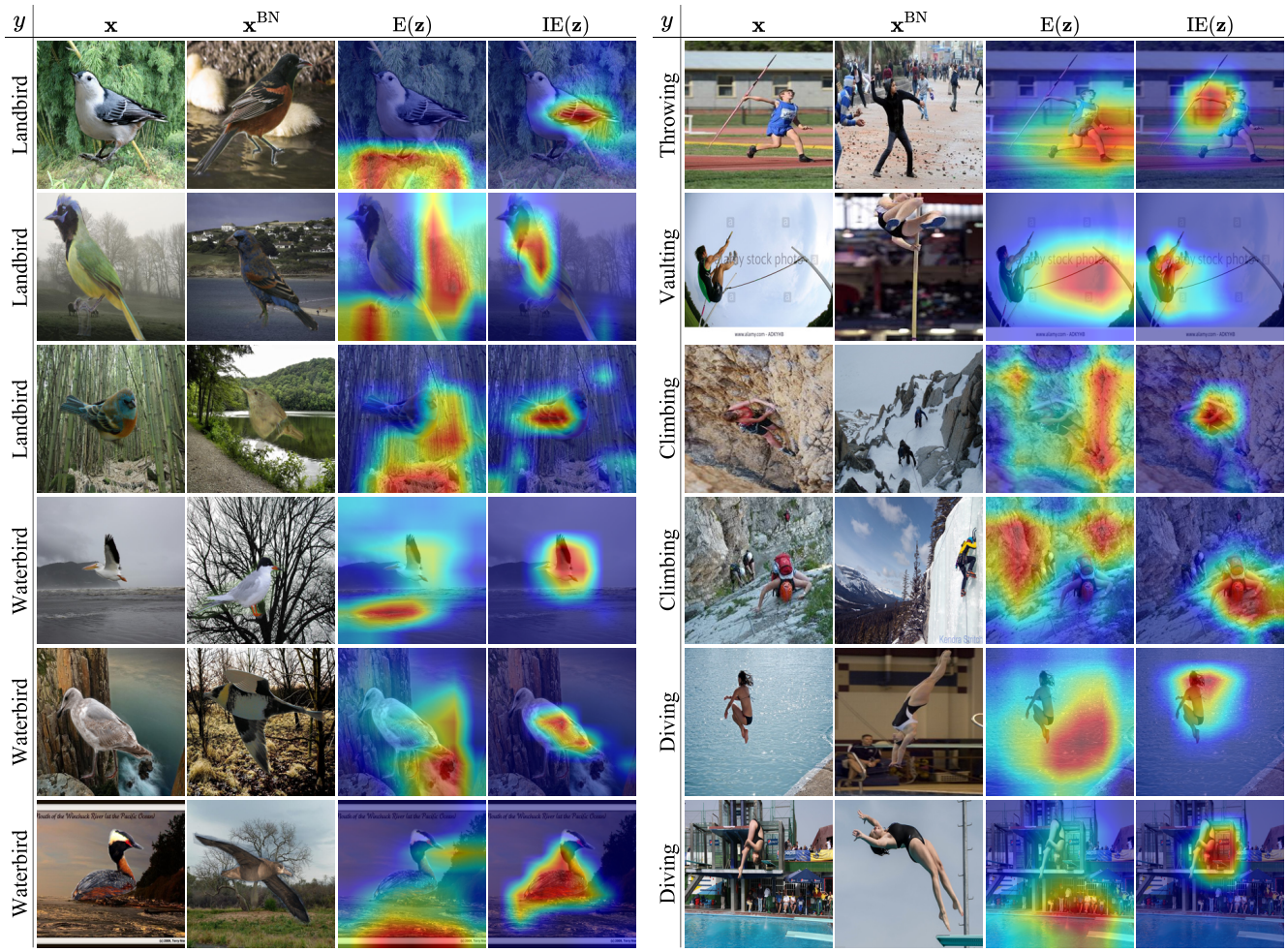
To further verify this, we present f_b's classification loss of samples in $\mathcal{D}^{\text{BN}}_{\text{cand}}$ during the training in Fig. 1. The BFFHQ dataset [6] with a bias severity of 1% is used for the analysis. In Fig. 1, the dotted lines denote the distribution of f_b's classification loss at the early stage of training (1K-th iteration), and the solid lines indicate that of the later stage of training (50K-th iteration). The results show that the f_b loss of BC samples (blue lines) largely increases at the later stage of training compared to the early stage, unlike BA samples (red lines). This demonstrates that the BN score as a loss weight can effectively upweight the training losses when BC samples are chosen as $\mathbf{x}^{\text{BN}}$.

B. Samples having negative BN score

Figure 2. The examples of samples that have negative BN scores at the later stage of training.

As mentioned in Sec. 3.2 in the main paper, we further filter out the samples with negative BN scores from $\mathcal{D}^{\text{BN}}_{\text{cand}}$ to mainly exploit the BC samples as $\mathbf{x}^{\text{BN}}$. Here, we expect that the samples with negative BN scores are mostly BA samples. To investigate the samples with negative BN scores, we chose the samples that were erroneously incorporated into $\mathcal{D}^{\text{BN}}_{\text{cand}}$ initially but excluded at the later stage of training (*i.e.*, 50K-th iteration), exhibiting negative BN scores. This process is repeated five times, and we visualize the samples chosen more than three times in Fig. 2. We use the BFFHQ dataset with a 1% bias severity for the experiment.

We observe that the samples with negative BN scores in $\mathcal{D}^{\text{BN}}_{\text{cand}}$ are mostly BA samples. As shown in the figure, while the samples obviously contain bias attributes (*i.e.*, features representing female or male), the samples mostly have extreme shade, blur, saturation, or unusual makeup, exhibiting non-typical appearance. Although the bias attributes



(a) Waterbirds (b) BAR

Figure 3. Additional visualization results of the spatial guidance using (a) Waterbirds and (b) BAR dataset. Given bias-contrastive pairs, $\mathbf{x}$ and $\mathbf{x}^{\text{BN}}$, $\text{E}(\mathbf{z})$ indicates the regions originally focused on by f_d and $\text{IE}(\mathbf{z})$ shows the regions highlighted by our IE weight.

are known to be easy to learn, the non-typical appearance prevents f_b from detecting such bias attributes in the early stage of training. In Sec. 4.5 of the main paper, we verify that employing such BA samples as $\mathbf{x}^{\text{BN}}$ largely degrades the debiasing performance by allowing the bias attributes to be included in the common features between $\mathbf{x}$ and $\mathbf{x}^{\text{BN}}$. Our BN score effectively alleviates this issue by filtering out such BA samples from $\mathcal{D}^{\text{BN}}_{\text{cand}}$.

C. Importance of BN sample selection

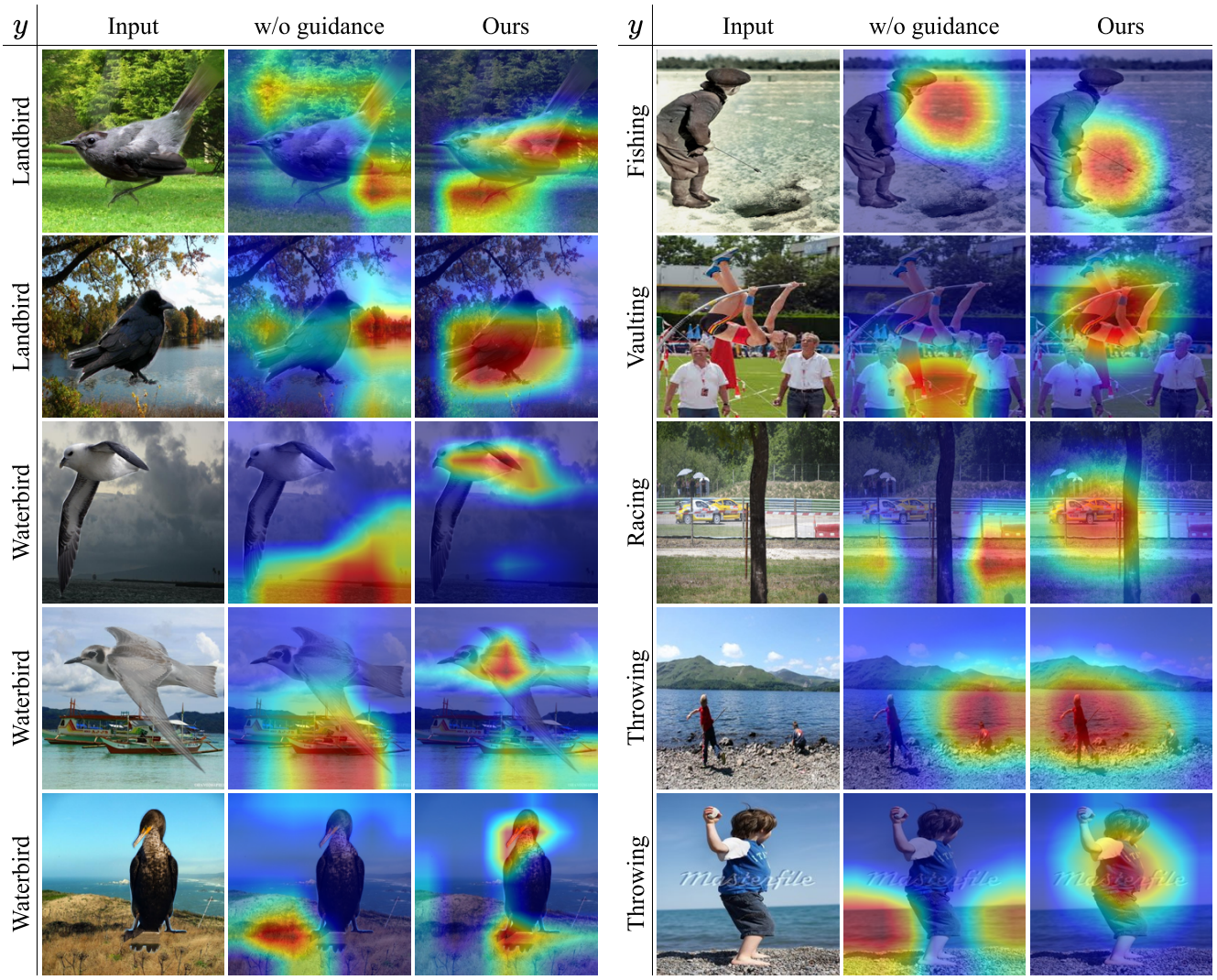
We analyze the effect of the BC sample ratio in $\mathcal{D}^{\text{BN}}$ on debiasing performance. We measure the accuracy using the BFFHQ dataset with a 1% bias severity by varying the number of BA and BC samples in $\mathcal{D}^{\text{BN}}$. Table 1 shows that higher accuracy is achieved for more BC samples and a lower ratio of BA to BC samples in $\mathcal{D}^{\text{BN}}$. Overall, our method constantly shows performance gain, except for the last column ({#BC in $\mathcal{D}^{\text{BN}}$/#BC in $\mathcal{D}$, #BA in $\mathcal{D}^{\text{BN}}$/#BC in $\mathcal{D}^{\text{BN}}$}-{1.0, 10.0}). It is crucial not to select too many BA samples as $\mathbf{x}^{\text{BN}}$.

#BC in $\mathcal{D}^{\text{BN}}$/#BC in $\mathcal{D}$	0.1	0.5	1.0	1.0	1.0	1.0	1.0
#BA in $\mathcal{D}^{\text{BN}}$/#BC in $\mathcal{D}^{\text{BN}}$	0.0	0.0	0.0	0.1	1.0	2.0	10.0
Accuracy	75.84	78.12	81.40	80.24	77.48	75.48	70.90

Table 1. Importance of $\mathbf{x}^{\text{BN}}$ selection.

D. Comparison to recent sample selection methods

Our BN score is designed to further filter out BA samples in $\mathcal{D}^{\text{BN}}_{\text{cand}}$, improving debiasing performance (Sec. 4.5 in the main paper). We compare BC sample selection in recent methods with ours using



(a) Waterbirds (b) BAR

Figure 4. Additional comparison of the region focused by a debiased model trained with and without our method. We compare Grad-CAM results on the test set of (a) Waterbirds and (b) BAR.

BFFHQ with a 1% bias severity. Let $\mathcal{S}$ be a set of samples identified as BC samples from training data $\mathcal{D}$. {#BC in $\mathcal{S}$/#BC in $\mathcal{D}$, #BA in $\mathcal{S}$/#BC in $\mathcal{S}$} is {75.63, 10.18}-BE [8], {27.29, 4.32}-DCWP [10], and {50.0, 0.89}-Ours, respectively. Our method has the least number of BA compared to BC samples in $\mathcal{S}$ while preserving half of the total BC samples.

E. Additional qualitative results

E.1. Visualization of the guidance during training

In addition to Fig. 3 of the main paper, we provide supplementary qualitative results that present the features that the current model f_d focuses on (*i.e.*, $\mathrm{E}(\mathbf{z})$) and the features emphasized by the guidance (*i.e.*, $\mathrm{IE}(\mathbf{z})$) during the training in Fig. 3. We use the Waterbirds and BAR datasets with a bias severity of 1% for the analysis. We train f_d during 10K iterations and obtain the visual explanation map $\mathrm{E}(\mathbf{z})$ for the ground-truth label using Grad-CAM [12]. The min-max normalization is applied to the values of $\mathrm{E}(\mathbf{z})$ and $\mathrm{IE}(\mathbf{z})$ for visualization. We intentionally select the BA sample and BC sample as $\mathbf{x}$ and $\mathbf{x}^{\mathrm{BN}}$, respectively, to compose a bias-negative pair.

As shown in Fig. 3, our IE weight (*i.e.*, $\mathrm{IE}(\mathbf{z})$) appropriately emphasizes the regions of the intrinsic features while $\mathrm{E}(\mathbf{z})$ shows that the current model f_d relies on the bias attributes for prediction. For example, in the Waterbirds dataset, $\mathrm{IE}(\mathbf{z})$ properly enhances the intrinsic features of a

bird such as wings, a body, or a neck, while $\mathrm{E}(\mathbf{z})$ highlights the background features such as the land, the forest or the water. Also, in the BAR dataset, $\mathrm{IE}(\mathbf{z})$ emphasizes the arm throwing the javelin, the motion of a person vaulting, climbing, or diving, while the current f_d mainly focuses on the biased features such as the playing field, the sky, the mountain, or the water. These results verify the validity of our IE weight $\mathrm{IE}(\mathbf{z})$ as guidance for emphasizing the intrinsic features in $\mathbf{x}$ that are under-exploited yet.

E.2. Effect of intrinsic feature guidance on debiasing

We present an additional qualitative analysis regarding the effectiveness of the intrinsic feature guidance to supplement Fig. 4 in the main paper. Fig. 4 illustrates the Grad-CAM [12] results of the model trained with and without our method. Here, the model trained without our method is the same as LfF+BE [8]. We train the models with the Waterbids and the BAR datasets with a bias severity of 1% and apply the Grad-CAM to the test samples for visualization. The highlighted regions indicate the features that the model mainly employs for prediction.

Fig. 4 (a) shows that our approach properly focuses on the intrinsic features of the bird (*e.g.*, wings, a beak, or feet), while the model trained without our guidance mostly concentrates on the bias features (*e.g.*, the water or trees). For the BAR dataset in Fig. 4 (b), our model principally exploits the action of a person (*e.g.*, fishing, vaulting, or throwing) or the racing car for prediction, while the model without our guidance focuses on the backgrounds (*e.g.*, the playing field or the water). The results demonstrate the effectiveness of our method in guiding the model to learn intrinsic features.

F. Additional quantitative results

F.1. Quantitative results with standard deviations

In Table 1 of our main paper, we report the quantitative comparison results with classification accuracies on the test set which are averaged across five independent experiments with different random seeds. We additionally provide the standard deviations of the classification accuracies in Table 2 and Table 3. Each table shows the results of the synthetic dataset (*i.e.*, Waterbirds) and the real-world dataset (*i.e.*, BFFHQ and BAR), respectively. Since the BAR dataset lacks explicit bias labels, approaches such as LNL and EnD that necessitate explicit bias labels are not applicable to the BAR dataset. The baseline results for the BFFHQ and the BAR dataset are from the results reported in BE [8] except for DCWP [10].

F.2. Comparison to recent baseline

Our primary contribution lies in providing the model with explicit spatial guidance for intrinsic features by examining features that commonly appear in bias-contrastive pairs. The intrinsic feature exists in generally appearing features within a class, however, this property has not been tackled to provide intrinsic feature guidance in prior studies to the best of our knowledge. While recent debiasing approaches aim to encourage the model to learn intrinsic features, they fail to directly indicate where the model should focus to learn the features.

For instance, MaskTune [1] expects the model to learn intrinsic features by fine-tuning the model with the data whose already-explored area is masked out using Grad-CAM. However, simply exploring the unmasked area cannot inform the model where exactly the intrinsic features are located. In this case, the model may rather focus on non-intrinsic features during the fine-tuning. We experiment on real-world datasets with a 1% bias severity: {58.00, 69.42}-MaskTune and {77.56, 75.14}-Ours for {BFFHQ, BAR}. Ours achieves better debiasing performance by providing explicit spatial guidance for intrinsic features based on common features in bias-contrastive pairs.

A recent pair-wise debiasing method $\mathcal{X}^2$-model [16] encourages the model to retain intra-class compactness using samples generated via feature-level interpolation between BC and BA samples. However, $\mathcal{X}^2$-model does not inform the model where the intrinsic features are located in the interpolated features. Simply making samples closer to the interpolated samples does not assure the model to focus on the intrinsic features. In contrast, our method directly encourages the model to focus on the area of the intrinsic features.

Also, we conduct a quantitative comparison to the recently proposed debiasing approach, DCWP [10], in Table 3. We use real-world datasets, BFFHQ and BAR, with various bias severity. For a fair comparison, we utilize ResNet18, which is the same architecture as ours. The ImageNet pretrained weight is employed only for the BAR dataset. The results demonstrate the superiority of our method over the DCWP, where ours provides the model with explicit guidance for intrinsic features for debiasing, unlike DCWP.

F.3. Worst accuracy between the accuracy of BA and BC samples in Waterbirds

To further analyze our model's performance on the Waterbirds dataset, we measure the accuracy of BA and BC samples separately, where the class accuracy values are averaged. Then, we report the worst accuracy between them in Table 4. The results show that ours achieves the highest worst accuracy compared to other baselines.

G. Detailed description of datasets

We utilize Waterbirds [11], BFFHQ [6], and BAR [9] dataset. First, the Waterbirds dataset is composed of two

Method	Waterbirds			
	0.5	1.0	2.0	5.0
Vanilla [3]	57.41 ±0.74	58.07 ±1.00	61.04 ±0.55	64.13 ±0.14
HEX [15]	57.88 ±0.83	58.28 ±0.67	61.02 ±0.48	64.32 ±0.62
LNL [5]	58.49 ±0.81	59.68 ±0.78	62.27 ±0.91	66.07 ±1.15
EnD [13]	58.47 ±0.97	57.81 ±1.04	61.26 ±0.54	64.11 ±0.52
ReBias [2]	55.44 ±0.24	55.93 ±0.66	58.53 ±0.52	62.14 ±1.03
LfF [9]	60.66 ±0.77	61.78 ±1.53	58.92 ±2.93	61.43 ±1.92
DisEnt [7]	59.59 ±1.67	60.05 ±0.82	59.76 ±1.26	64.01 ±1.36
LfF+BE [8]	61.22 ±2.54	62.58 ±1.12	63.00 ±1.18	63.48 ±0.56
DisEnt+BE [8]	51.65 ±1.45	54.10 ±1.04	53.43 ±1.42	54.21 ±1.36
Ours	**63.64** ±1.63	**65.22** ±0.95	**65.23** ±1.06	**66.33** ±1.42

Table 2. Comparison to the baselines. We measure the classification accuracy on test sets of the Waterbirds dataset with different bias severities. The best accuracy values are in bold. Results with standard deviations are provided in the Supplementary.

Method	BFFHQ				BAR	
	0.5	1.0	2.0	5.0	1.0	5.0
Vanilla [3]	55.64 ±0.44	60.96 ±1.00	69.00 ±0.50	82.88 ±0.49	70.55 ±0.87	82.53 ±1.08
HEX [15]	56.96 ±0.62	62.32 ±1.21	70.72 ±0.89	83.40 ±0.34	70.48 ±1.74	81.20 ±0.68
LNL [5]	56.88 ±1.13	62.64 ±0.99	69.80 ±1.03	83.08 ±0.93	-	-
EnD [13]	55.96 ±0.91	60.88 ±1.17	69.72 ±1.14	82.88 ±0.74	-	-
ReBias [2]	55.76 ±1.50	60.68 ±1.24	69.60 ±1.33	82.64 ±0.64	73.04 ±1.04	83.90 ±0.82
LfF [9]	65.19 ±3.23	69.24 ±2.07	73.08 ±2.70	79.80 ±1.09	70.16 ±0.77	82.95 ±0.27
DisEnt [7]	62.08 ±3.89	66.00 ±1.33	69.92 ±2.72	80.68 ±0.25	70.33 ±0.19	83.13 ±0.46
LfF+BE [8]	67.36 ±3.10	75.08 ±2.29	80.32 ±2.07	85.48 ±2.88	73.36 ±0.97	83.87 ±0.82
DisEnt+BE [8]	67.56 ±2.11	73.48 ±2.12	79.48 ±1.80	84.84 ±2.11	73.29 ±0.41	84.96 ±0.69
DCWP [10]	64.08 ±1.08	67.44 ±2.87	75.24 ±1.73	85.00 ±0.94	69.63 ±0.85	81.89 ±0.68
Ours	**71.68** ±1.74	**77.56** ±1.24	**83.08** ±1.69	**87.60** ±1.68	**75.14** ±0.82	**85.03** ±0.64

Table 3. Comparison to the baselines. We measure the classification accuracy on test sets of the BFFHQ and BAR datasets with different bias severities. The best accuracy values are in bold. The hyphen mark (-) means it is not applicable. Results with standard deviations are provided in the Supplementary.

classes of bird images and has background bias. In the training set, the waterbirds are mostly with the water background and the landbirds are with the land background. The number of BA samples and that of BC samples are balanced in the test set. By following Sagawa *et al.* [11], we utilize two datsets, the Caltech-UCSD Birds-200-2011 (CUB) dataset [14] and the Places[1] dataset [18], to construct the Waterbirds dataset. The bird images are segmented from the CUB dataset, and the segmented birds are combined with the background images from the Place dataset. We employ the code released by Sagawa *et al.* [11][2] for constructing the dataset. As mentioned in the repository, a few landbirds (Eastern Towhees, Western Meadowlarks, and Western Wood Pewees) in the original dataset are incorrectly labeled as waterbirds. Therefore, we correct their labels to landbirds for the experiments.

The BFFHQ is initially presented by Kim *et al.* [6] and constructed by modifying the FFHQ dataset [3]. In the BFFHQ, the bias attribute is the gender and the intrinsic attribute is the age. Specifically, most of the young people are female, and most of the old people are male in the training dataset.

Lastly, the BAR dataset is introduced by Nam *et al.* [9]. The dataset contains six action classes (*i.e.*, Climbing, Diving, Fishing, Vaulting, Racing, Throwing) and each class is biased to a certain background (*i.e.*, RockWall, Underwater, WaterSurface, Sky, APavedTrack, PlayingField). In the test

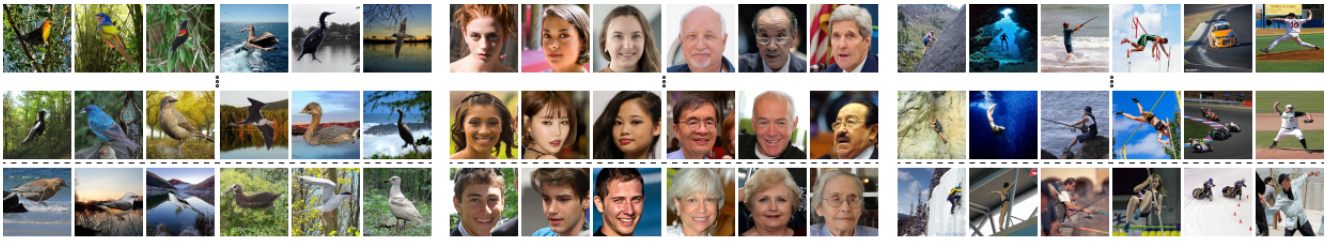
[1] CC BY

[2] https://github.com/kohpangwei/group_DRO

[3] BY-NC-SA 4.0

BS	Vanilla [3]	HEX [15]	LNL [5]	EnD [13]	ReBias [2]	LfF [9]	DisEnt [7]	LfF+BE [8]	DisEnt+BE [8]	Ours
0.5	24.08 ±1.56	28.20 ±3.07	26.08 ±1.64	28.29 ±3.53	27.00 ±1.10	56.22 ±6.07	38.07 ±11.01	55.15 ±2.78	36.60 ±10.88	59.12 ±3.67
1.0	24.78 ±2.45	26.32 ±2.90	29.72 ±3.45	25.69 ±2.41	27.95 ±1.56	59.07 ±3.40	47.02 ±7.26	55.53 ±1.60	28.35 ±4.17	63.05 ±1.97
2.0	34.39 ±2.24	32.12 ±2.89	33.92 ±1.94	32.94 ±1.48	32.16 ±0.76	53.07 ±6.74	44.93 ±8.54	52.91 ±2.62	31.08 ±6.01	61.71 ±4.94
5.0	38.34 ±1.05	39.08 ±0.92	43.22 ±1.94	40.91 ±1.11	39.72 ±1.11	58.05 ±2.37	52.96 ±6.33	48.48 ±3.72	37.92 ±6.47	58.60 ±3.32

Table 4. The worst accuracy between the accuracy of BA and BC samples in the Waterbirds dataset. BS is bias severity.



(a) Waterbirds (b) BFFHQ (c) BAR

Figure 5. Visualization of datasets used in the experiments. A group of three columns represents each class for (a) Waterbirds-{Landbird, Waterbird} and (b) BFFHQ-{Young, Old}, and each column of (c) BAR-{Climbing, Diving, Fishing, Vaulting, Racing, Throwing} represents a distinct class. The samples above the dashed line are bias-aligned samples and the below ones are bias-conflicting samples.

set, such correlations do not exist. For the experiments, we use the BFFHQ dataset and BAR dataset released by Lee *et al.* [8][4]. The examples of the BA samples and BC samples in each dataset are shown in Fig. 5.

H. Implementation details

Following the previous studies [7–9], we utilize ResNet18 [3] for the biased model f_b and the debiased model f_d. Also, f_d^{emb} indicates the subnetwork before the average pooling layer, and f_d^{cls} consists of an average pooling layer and a linear classifier that outputs logits, where $f_d(\mathbf{x}) = f_d^{\text{cls}}\left(f_d^{\text{emb}}(\mathbf{x})\right)$. Before training, f_b and f_d are initialized with the ImageNet pretrained weight for the BAR dataset, while we randomly initialize the models for the other datasets. This is because the size of the BAR dataset is extremely small compared to the others [8].

During training f_d, we employ the sample reweighting value $w(\mathbf{x})$ termed as relative difficulty score [9], as mentioned in Sec. 3.4 in the main paper. $w(\mathbf{x})$ is calculated as follows:

$$w(\mathbf{x}) = \frac{\mathcal{L}_{\text{CE}}(f_b(\mathbf{x}), y)}{\mathcal{L}_{\text{CE}}(f_b(\mathbf{x}), y) + \mathcal{L}_{\text{CE}}(f_d(\mathbf{x}), y)}. \quad (1)$$

This score assigns a high weight to the BC samples and a low weight to the BA samples. This encourages f_d to mainly learn intrinsic features by emphasizing BC samples with $w(\mathbf{x})$.

[4]https://github.com/kakaoenterprise/BiasEnsemble

The models are trained for 50K iterations with a batch size of 64. The horizontal flip and a random crop with a size of 224 are used for data augmentation during the training. All the models are trained with the Adam optimizer. The learning rate is set as 1e-4 for the Waterbirds and the BFFHQ dataset, and 1e-5 for the BAR dataset. The hyperparameters of α_l, α_s, and τ are set as 0.1, 0.9, and 2, respectively, for all the datasets. We apply class-wise max normalization to our BN score to consider the different ranges of the scores across the classes for stability of training.

During the training, we aim to select an auxiliary sample that has no bias attribute but has the same class label with $\mathbf{x}$ as $\mathbf{x}^{\text{BN}}$ from $\mathcal{D}^{\text{BN}}$. If no sample in $\mathcal{D}^{\text{BN}}$ has the same label as $\mathbf{x}$, we select the sample that has the same label with $\mathbf{x}$ from $\mathcal{D}^{\text{BN}}_{\text{cand}}$. In a case where there's no sample with the same label as $\mathbf{x}$ in both $\mathcal{D}^{\text{BN}}$ and $\mathcal{D}^{\text{BN}}_{\text{cand}}$, we sample $\mathbf{x}^{\text{BN}}$ that has the same label with $\mathbf{x}$ from $\mathcal{D}$.

As described in Sec. 3.1 in the main paper, we utilize the pretrained biased models to construct $\mathcal{D}^{\text{A}}$, following the previous work [8]. We utilize ResNet18 [3] for the pretrained biased models, and all the pretrained biased models are randomly initialized before training. The models are trained for 1K iterations with the generalized cross entropy (GCE) loss [17]. Within each model, the samples with a high ground-truth probability (*i.e.,* ≥ 0.99) are considered as BA samples. Based on majority voting, we collect the samples that are considered as the BA sample by the majority of the models and construct the bias-amplified dataset $\mathcal{D}^{\text{A}}$. We use five pretrained biased models following the

study of Lee *et al*. [8]. Lee *et al*. demonstrate that adopting the additional biased models requires a negligible amount of additional computational costs and memory space. Note that the same biased models are utilized when constructing $\mathcal{D}^{\text{BN}}_{\text{cand}}$.

I. Limitations and future work

Although our BN score effectively encourages BC samples to be mainly adopted as auxiliary inputs, the auxiliary inputs still can include a few BA samples, as shown in Table 2 of Sec. 4.3 in the main paper. Accordingly, such BA samples may interfere with the model to capture the intrinsic features. Identifying intrinsic attributes without relying on auxiliary inputs can be one promising future work.

In addition, since our IE weight is designed to enhance intrinsic features by imposing spatially different values on the features, our method might be more effective especially when bias attributes are located in different regions from the intrinsic attributes. In this regard, applying channel-wise re-weighting [4] to our approach will further improve the general applicability of our method.

Despite the limitations above, we believe that our work poses the importance of enhancing intrinsic attributes for debiasing.

References

[1] Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, pages 23284–23296, 2022. 4

[2] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *Proc. the International Conference on Machine Learning (ICML)*, 2020. 5, 6

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016. 5, 6

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 7

[5] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019. 5, 6

[6] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proc. of the IEEE international conference on computer vision (ICCV)*, pages 14992–15001, 2021. 1, 4, 5

[7] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5, 6

[8] Jungsoo Lee, Jeonghoon Park, Daeyoung Kim, Juyoung Lee, Edward Choi, and Jaegul Choo. Revisiting the importance of amplifying bias for debiasing. In *Proc. the AAAI Conference on Artificial Intelligence (AAAI)*, pages 14974–14981, 2023. 3, 4, 5, 6, 7

[9] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4, 5, 6

[10] Geon Yeong Park, Sangmin Lee, Sang Wan Lee, and Jong Chul Ye. Training debiased subnetworks with contrastive weight pruning. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 7929–7938, 2023. 3, 4, 5

[11] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *Proc. the International Conference on Learning Representations (ICLR)*, 2020. 4, 5

[12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. of the IEEE international conference on computer vision (ICCV)*, 2017. 3, 4

[13] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 13508–13517, 2021. 5, 6

[14] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[15] Haohan Wang, Zexue He, Zachary L. Lipton, and Eric P. Xing. Learning robust representations by projecting superficial statistics out. In *Proc. the International Conference on Learning Representations (ICLR)*, 2019. 5, 6

[16] Yi-Kai Zhang, Qi-Wei Wang, De-Chuan Zhan, and Han-Jia Ye. Learning debiased representations via conditional attribute interpolation. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 7599–7608, 2023. 4

[17] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv preprint arXiv:1805.07836*, 2018. 6

[18] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(6):1452–1464, 2017. 5