A^2: Adaptive Augmentation for Effectively Mitigating Dataset Bias

Jaeju An[1], Taejune Kim[1], Donggeun Ko[4], Sangyup Lee[1], and Simon S. Woo[1,2,3,⋆]

[1] Dept. of Computer Science & Engineering, Sungkyunkwan University, South Korea
[2] Dept. of Artificial Intelligence, Sungkyunkwan University, South Korea
[3] Dept. of Applied Artificial Intelligence, Sungkyunkwan University, South Korea
[4] Dept. of Applied Data Science, Sungkyunkwan University, South Korea
{anjaeju,taemo,seanko,sangyup.lee,swoo}@g.skku.edu

Abstract. Recently, deep neural networks (DNNs) have become the de facto standard to achieve outstanding performances and demonstrate significant impact on various computer vision tasks for real-world scenarios. However, the trained networks can often suffer from overfitting issues due to the unintended bias in a dataset causing inaccurate, unreliable, and untrustworthy results. Thus, recent studies have attempted to remove bias by augmenting the bias-conflict samples to address this challenge. Yet, it still remains a challenge since generating bias-conflict samples without human supervision is generally difficult. To tackle this problem, we propose a novel augmentation framework, Adaptive Augmentation (A^2), based on a generative model that help classifiers learn debiased representations. Our framework consists of three steps: 1) extracting bias-conflict samples from a biased dataset in an unsupervised manner, 2) training a generative model with the biased dataset and adapting the learned biased distribution to the extracted bias-conflict samples' distribution, and 3) augmenting bias-conflict samples by translating bias-align samples. Therefore, our classifier can effectively learn the debiased representation without human supervision. Our extensive experimental results demonstrate that A^2 effectively augments bias-conflict samples, mitigating widespread bias issues. The code is available in here[5].

Keywords: Computer Vision · Debiasing · Image Translation.

1 Introduction

Recently, deep neural networks (DNNs) have achieved great success across various research fields, including image classification [21], object detection [34], semantic segmentation [22], and even image generation [8]. Nonetheless, overfitting, a well-known problem in DNNs, causes the models to produce inaccurate

⋆ Corresponding author
[5] https://github.com/anjaeju/A2-Adaptive-Augmentation-for-Effectively-Mitigating-Dataset-Bias

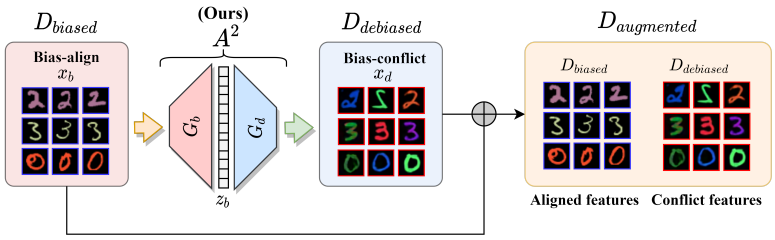

Fig. 1: Illustration of our A^2 framework. A^2 projects bias-align samples x_b, and outputs translated bias-conflict samples x_d. The classifier trained with D_{biased} produces 33.36%, while the classifier trained with $D_{augmented}$ produces 67.47% accuracy on bias-conflict samples in Colored MNIST dataset.

and unreliable results leading to failure in making proper decisions [31]. This phenomenon is closely related to "dataset bias," where unintended bias exists in the training dataset. In particular, unintended bias indicates that a large number of samples appear with similar task-irrelevant features in a visual context.

Consider the case where we need to classify frog images with different backgrounds. As such, frogs and the background is assigned as *task-relevant feature* and *task-irrelevant feature*, respectively. Most of the frogs could be located in a green pond background (*bias-align*), while a few could be positioned in a swamp or an asphalt road against the green pond (*bias-conflict*). Since the models tend to learn task-irrelevant features (easy-to-learn) as a cue for the labels [19], the models fail to properly learn the task-relevant features, raising questions on the actual performance in the presence of different biases.

To handle such aforementioned bias issue, numerous research has been proposed to 'debias' the dataset bias by defining specific bias types [16, 20, 27] or learning debiased representation without explicitly defining bias types [2, 30, 3, 5, 6]. Recent advancements in debiasing have demonstrated that augmentation is one of the most promising approaches for mitigating dataset bias [19]. However, augmenting bias-conflict samples without human supervision remains challenging due to the complex properties of different biases, such as texture or shape [19]. Therefore, given the difficulties, we aim to build a practical debiasing method that generates bias-conflict samples without any prior knowledge.

In this paper, we first conduct a set of preliminary experiments to illustrate the importance of augmenting bias-conflict samples, when applying augmentation in biased settings. We find that augmenting only bias-conflict samples significantly improves the classification performance. Based on this finding, we propose a novel augmentation framework, Adaptive Augmentation (A^2), for augmenting bias-conflict samples in an unsupervised manner. Figure 1 illustrates our approach, an end-to-end debiasing pipeline to prevent dataset bias by translating bias-align samples into bias-conflict samples which effectively increases the number of bias-conflict samples to prevent overfitting to task-irrelevant features.

Note that our proposed framework can also be seen as a pre-processing method that effectively mitigates dataset bias.

Adaptive Augmentation (A^2) consists of three main components: 1) extracting a few numbers of bias-conflict samples from any biased dataset without human supervision, 2) training a biased generative model with bias-align samples and adapting the learned model's parameters to the extracted samples' distribution, and 3) augmenting bias-conflict samples by translating bias-align samples with the trained generative models. These augmented images contribute to learning task-relevant features for a biased classifier. We demonstrate that our A^2 outperforms the baselines in comprehensive debiasing benchmark datasets through extensive experiments. Moreover, we confirm that our method performs effectively in an extremely biased setting, where we have very few bias-conflict samples in each dataset. The contributions of our work are summarized as follows:

- We propose a novel debiasing augmentation framework, A^2, which leverages an unsupervised algorithm for extracting bias-conflict samples and exploits few-shot adaptation by adjusting the distribution of the biased generative model to bias-conflict distribution.
- We evaluate the performance of our A^2 through quantitative and qualitative analysis for both synthetic and real-world datasets. We demonstrate that our method achieves the state-of-the-art performance in biased settings.
- We investigate the reason for performance improvement through comprehensive and carefully constructed ablation studies. We believe that our approach has a broader impact by presenting a new application of generative models to solve challenging bias issues for a variety of computer vision applications.

2 Related Work

2.1 Benchmark Datasets for Debiasing

Recently, synthetic or real-world datasets have been created and released publicly to foster the debiasing research field, as shown in Fig. 2.

First, Colored MNIST (CMNIST) and Corrupted CIFAR10 (CCIFAR10) are synthetic datasets built by manually injecting distinct biases into existing MNIST and CIFAR10 datasets. CMNIST dataset is created by adding a color bias to the MNIST [18, 19] dataset, as shown in Fig. 2a. Instead of adding color bias like CMNIST, CCIFAR10 is constructed by applying ten distinct noise corruption to each of the labels in CIFAR10 [17, 12], as depicted in Fig. 2b. While these two synthetic datasets have been extensively used in previous studies, the challenge for mitigating real-world bias remains, as synthetic datasets are relatively simple to cover real-world bias, such as age or gender.

Second, Biased FFHQ (BFFHQ) and Biased Action Recognition (BAR) datasets are released to mitigate biases in real-world data. BFFHQ [19] is curated from the FFHQ [14] dataset, which contains high-quality images of human faces. From FFHQ, the BFFHQ dataset selects age as a task-relevant feature and gender as a task-irrelevant feature. Accordingly, the majority of the young

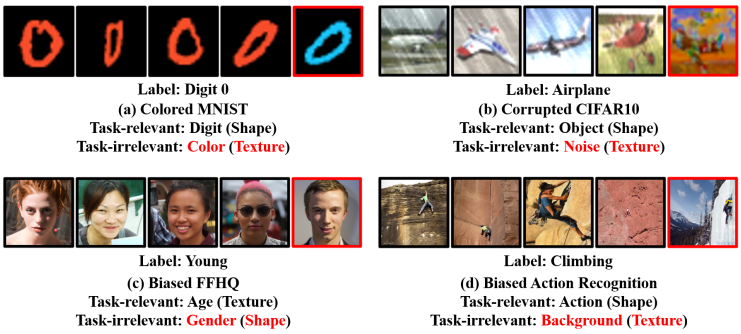

Fig. 2: Sample images from benchmark datasets. We provide task-relevant and task-irrelevant features in each dataset, where each class is represented by a group of images comprising four bias-align samples and a bias-conflict sample.

are women as shown in Fig. 2c. On the other hand, the BAR [23] consists of real-world images of six action classes in distinct places, assuming that the classifier is likely to be biased in background (texture) bias. For instance, as shown in Fig. 2d, most bias-align samples from the *climbing* label contain a climber climbing a rock cliff. In contrast, the bias-conflict sample exhibits a climber with a glacier cliff. Unlike synthetic datasets containing texture biases, real-world datasets have both texture and shape biases which are relatively difficult to handle.

2.2 Existing Methods for Debiasing

Numerous previous studies have been performed to mitigate dataset biases. We mainly investigate existing debiasing methods in the context of augmentation: debiasing without data augmentation and with augmentation.

Debiasing without Augmentation. Debasing can be performed through an explicit definition of the bias type [16, 20, 27, 7, 23, 30]. Li and Vasconcelos et al. [16] and Kim et al. [20] demonstrated that a specific color bias could be relieved by normalizing the biased distribution or utilizing biased RGB values as a clue for classification. Sagawa et al. [27] proposed the groupDRO, a debiasing method which clusters subgroups in the dataset with explicit supervision. GroupDRO is then further improved with various modifications [7, 32, 26]. Moreover, Kim et al. [16] employed a regularization loss to inhibit the model from learning unwanted bias in the dataset. However, none of the above are applicable in real-world datasets since defining whole bias types is unfeasible when dealing with large-scale real-world datasets [19]. In contrast, recent approaches widely adopt the debiasing method without defining prior knowledge of the bias. Nam et al. [23] proposed LfF, a state-of-the-art method for simultaneously training biased and debiased models to amplify the influence of bias-align samples on biased models and bias-conflict samples on debiased models, respectively.

Debiasing with Augmentation. Debiasing with augmentation is less explored compared to the debiasing approaches without augmentation mentioned above. Augmentation can be conducted on the image level to mitigate shape-texture biases. Agarwal et al. [1] mitigated textural bias by training on the 'Styled ImageNet' [6, 13] dataset, which contains severely distorted textures of the ImageNet data. Furthermore, MixStyle, proposed by Zhou et al. [33], encouraged the classifier to extract more generalized features against the texture bias by shuffling feature-level statistics. These techniques, however, have a drawback in that they are limited to dealing with textual bias.

Recently, Lee et al. [19] presented the DisEnt approach, which augments bias-conflict samples by disentangling task-relevant and task-irrelevant features and permuting each bias feature vector to the other task-relevant feature vectors. Motivated by this observation, we design an augmentation framework that adapts task-irrelevant features regardless of the bias type to address both texture and shape biases without supervision, enhancing the performance of classification methods.

3 Importance of Augmenting Bias-conflict Samples

3.1 Overview

Data augmentation is a crucial way to boost the performance of DNNs, which applies various transformations to the original data, and compensates for the lack of datasets [29, 28]. Various augmentation techniques are available and have been shown to improve the model performance; however, when applied to biased datasets, augmentation might amplify the bias in the dataset by increasing bias-align samples. Therefore, we conduct the following experiments assuming that indiscriminate augmentation can degrade the classification performance:

- Case 1: Augmenting only bias-align samples.
- Case 2: Augmenting only bias-conflict samples.
- Case 3: Augmenting both bias-align and conflict samples.

We can find that Case 1 and 2 augment bias-align and bias-conflict samples, respectively, whereas Case 3 augments both bias-align and bias-conflict samples. These case studies can validate the degree of bias amplification according to the augmented sample type. The above experiments show the effect of augmentation methods for each case while using only the simple augmentations (i.e., the random crop and rotation) that do not affect the image's texture.

Dataset and Classifier. We demonstrate our method's performance on two synthetic and one real-world datasets, CMNIST, CCIFAR10, and BFFHQ. For CMNIST and CCIFAR10, we use bias ratios 99.5% and 95%, respectively. The model's performance is evaluated using a test set composed solely of bias-conflict samples. See Appendix A.1 for more details.

Results. Table 1 summarizes the model's performance in each case mentioned above. The *baseline* column indicates how well the vanilla model performs on

Table 1: Performance comparison on the bias-conflict test sets. Each case indicates different augmentation scenarios. We observe performance degradation in Cases 1 and 3 but performance improvement in Case 2. It denotes that applying data augmentation in a biased setting can cause the bias to exacerbate. We report the average accuracy over three runs. Bold indicates the best accuracy.

Dataset	Bias Ratio	Baseline	Case 1	Case 2	Case 3
Colored MNIST	99.5	33.36	24.89	**37.07**	33.76
	95	83.88	82.70	**84.10**	79.96
Corrupted CIFAR10	99.5	13.23	12.77	**13.31**	12.46
	95	27.37	27.27	**28.08**	27.60
Biased FFHQ	99.5	43.93	42.20	**45.67**	45.13

each test set. The model's performance in Case 1 (augmenting bias-align samples) declines by up to 8%, indicating that the dataset's bias has been amplified due to the augmented bias-align samples. In contrast, Case 2 (augmenting bias-conflict samples) illustrates that the model's performance increases by up to 4%, demonstrating that the dataset's bias has diminished. However, in Case 3 (augmenting both bias-align and bias-conflict samples), the model's performance remains close to the vanilla model without making any meaningful improvements. Therefore, it is desirable to augment only bias-conflict samples in biased settings. These findings established the validity of our hypothesis. With this motivation, we propose A^2 to augment bias-conflict samples effectively.

4 Design of A^2 framework

This section describes a novel augmentation framework, Adaptive Augmentation (A^2). First, we introduce our extraction method that selects a few bias-conflict samples from a biased dataset in an unsupervised manner. Second, we describe the detailed design of A^2, which utilizes a generative model and few-shot adaptation. Finally, we explain the training scheme to build a debiased classifier with an augmented dataset. We present the pipeline of our approach in Fig. 3.

4.1 Extracting Bias-conflict Samples

Suppose we can train a classifier to be biased regardless of the bias type (e.g., different bias types presented in Section 2.1). Then, the biased classifier predicts a target label with high confidence for unseen bias-align samples because the representation of the biased sample is similar to the trained biased samples [19, 2]. On the contrary, the classifier predicts with a low confidence for unseen bias-conflict sample, because its representation is different from the bias-align samples. This result leads to a higher cross-entropy loss for bias-conflict samples compared to the bias-align samples. In that case, we can easily identify bias-conflict samples with the higher cross-entropy loss [19, 2]. Therefore, we need

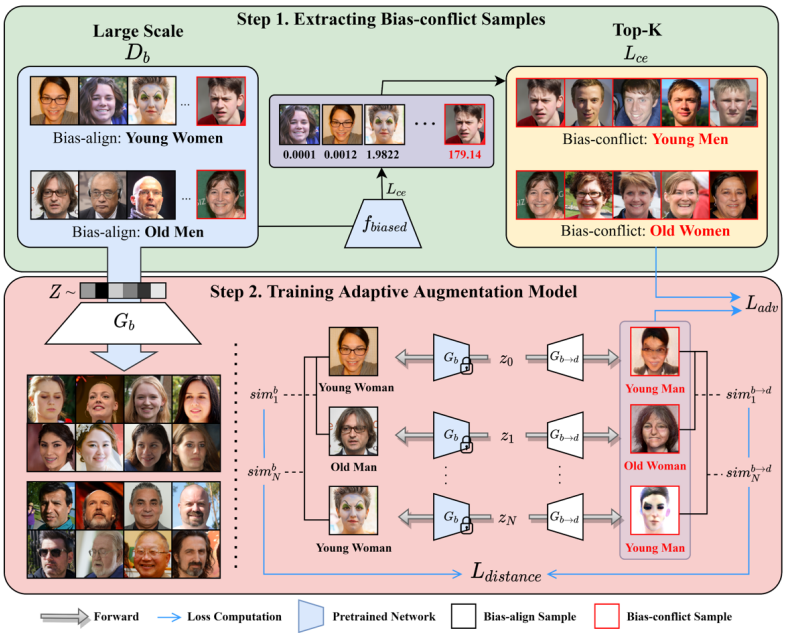

Fig. 3: Overview of our proposed debiasing method, Adaptive Augmentation (A^2) model. Our A^2 consists of two steps: 1) extracting bias-conflict samples and 2) training the Adaptive Augmentation model. In Step 1, we extract bias-conflict samples by sorting top-k L_{ce} values calculated by the biased classifier f_{biased}. Next, we adapt a biased generative model to the extracted bias-conflict distribution by minimizing L_{adv} and $L_{distance}$.

an unsupervised method to train a classifier to become biased without human supervision or pre-defined bias types.

To enable a classifier to be biased in an unsupervised manner, we need to keep emphasizing the impact of bias-align samples during training. Therefore, we employ generalized cross-entropy (GCE) loss [23] that amplifies the bias of the neural network without human-supervision. The equation of GCE loss is defined as follows:

$$GCE(p(x;\theta),y) = \frac{1-p_y(x;\theta)^q}{q}, \tag{1}$$

$$\frac{\partial GCE(p(x;\theta),y)}{\partial \theta} = p_y(x;\theta)^q \cdot \frac{\partial CE(p_y(x;\theta),y)}{\partial \theta}, \tag{2}$$

where $p(x;\theta)$ indicates the softmax output of a classifier θ, $p_y(x;\theta)$ represents the probability assigned to the target variable y with $q \in (0,1]$ that is a hyperparameter controlling the degree of amplification. The gradient of GCE loss emphasizes the gradient of CE loss by multiplying the probability p_y (i.e., confidence score). Since the task-irrelevant features are easy-to-learn [23] in the

early stages of training, the classifier first learns biased representations from the bias-align samples, producing higher confidence in the bias-align samples than that of bias-conflict samples. Thus, classifiers trained with GCE loss learn task-irrelevant features from biased images regardless of the type of bias each dataset has, ensuring that the classifiers become biased toward the aligned task-irrelevant features. With this assumption, we propose simple, yet powerful bias-conflict sample extraction algorithm, depicted in Step 1 of Fig. 3. Our algorithm is described as follows: we train our classifier with GCE loss and calculate CE loss for all samples in the training dataset to sort the loss values. After training, we extract the top-k samples as bias-conflict samples for utilizing them when adapting a generative model.

4.2 Training Generative Models

We elaborate on the training procedure of the bias adaptive generator that translates bias-align samples into bias-conflict samples, as described in Step 2 of Fig. 3. First, we need a pretrained generative model G_b trained on a biased dataset D_b. For G_b, we adopt a powerful generative model, StyleGAN2 [15], as our backbone for generating images (Note that any generative model can be used in this procedure, where we used StyleGAN2 since it is one of the high-performing GAN methods). G_b learns the distribution of D_b, by mapping 512 low-dimensional noise vector $z \sim p_z(z)$ to biased image $x_b \sim D_b$. The biased generative model can be obtained by using a GAN training procedure, with a discriminator D_m as follows:

$$L_{adv}(G_b, D_m) = D_m(G_b(z)) - D_m(x_b), \tag{3}$$

$$G_b^* = \arg\min_{G_b} \max_{D_m} \mathbb{E}_{z\sim p_z(z),\ x_b\sim D_b} L_{adv}(G_b, D_m), \tag{4}$$

where G_b^* denotes optimal weights of the model and the detailed training procedure is explained by Karras et al. [15]. After training the biased distribution, our G_b can generate samples that have bias-aligned representation, depicted in the left side of Step 2 in Fig. 3. Second, we convert biased generator (G_b) to debiased generator (G_d) by adapting the learned distribution to the bias-conflicting distribution with the extracted bias-conflict samples and all of the bias-align samples. In this case, we need to preserve the learned relationship between the generated samples from G_b and G_d during adaptation, as maintaining relative pairwise distances prevents mode collapse [25, 4, 10, 24]. To this end, we leverage the state-of-the-art few-shot adaptation method [24]. We initialize G_b and G_d with the pretrained weight of G_b, and sample N noise vectors $\{z_n\}_0^N$. Then, we forward z_i into the networks for obtaining the i^{th} generated samples $G_b(z_i)$ and $G_d(z_i)$. To match the relationship of generated samples from each network, we need to represent the relationship as probability distribution and minimize the distance of the two distributions. Therefore, we convert the generated samples

into probability distributions using cosine similarity as follows:

$$y_i^{b,l} = Softmax(\{sim(G_b^l(z_i), G_b^l(z_j))\}_{\forall j \neq i}), \tag{5}$$

$$y_i^{d,l} = Softmax(\{sim(G_d^l(z_i), G_d^l(z_j))\}_{\forall j \neq i}), \tag{6}$$

where l and sim indicates the l^{th} activation layers and cosine similarity, respectively. Now, we can preserve the relationship of the generated samples by minimizing $y_i^{b,l}$ and $y_i^{d,l}$ through KL-divergence:

$$L_{distance}(G_d, G_b) = \sum_{l,i} D_{KL}(y_i^{d,l} || y_i^{b,l}), \tag{7}$$

where the generated bias-align and bias-conflict samples have similar distribution. Thus, our final objective consists of two terms as follows:

$$G_d^* = \arg\min_{G_d} \max_{D_m} \mathbb{E}_{z \sim p_z(z),\ x_d \sim D_d} L_{adv}(G_d, D_m) + \lambda L_{distance}(G_d, G_b), \tag{8}$$

where D_d denotes the extracted bias-conflict samples, $L_{distance}$ for preserving the learned relationship, and L_{adv} for converting the learned distribution into another distribution.

4.3 Learning Debiased Representation with Adaptive Augmentation

After G_b and G_d converges, we generate debiased representation of bias-align sample x_d through an image-to-image translation approach as shown in the Fig. 1. Our A^2 takes the bias-align sample as an input, and projects it into the latent space (i.e., sample to latent), forming the biased latent vector z_b using G_b. The z_b contains label and bias information. Then, we forward z_b into G_d to generate bias-conflict representation of the input image (i.e., latent to sample). By keeping the label information for each x_b from the projected latent vector z_b, we do not need to label the generated samples manually (see Appendix A.2 for more details). After translating all x_b to debiased sample x_d, we obtain the debiased dataset D_d. By adding D_d to the original dataset, we have $D_{augmented} = D_b \cup D_d$ that supports to learn task-relevant features more effectively than training solely D_b, as task-irrelevant features are not aligned. Finally, we train the debiased classifier $f_{debiased}$ with $D_{augmented}$ using CE loss as follows:

$$f_{debiased}^* = \arg\min_f \frac{1}{N} \sum_{i=1}^{N} CE(f(x), y), \tag{9}$$

where we use $f_{debiased}^*$ for predicting bias-align and bias-conflict samples.

5 Experiments

5.1 Experimental Setup

Datasets. We use both synthetic (Colored MNIST and Corrupted CIFAR10) and real-world (Biased FFHQ and Biased Action Recognition) datasets. We

report bias-align and bias-conflict classification accuracy (%) on each test set, qualitative analysis on augmented samples, and ablation studies for our A^2. We further assess our A^2 with one-shot scenario, where only one bias-conflict sample per class exists, for demonstrating the validity of the proposed method in extremely biased settings; in this case, we adopt Colored MNIST, Corrupted CIFAR10, and Biased FFHQ datasets. See Appendix A.1 for more details (e.g., the number of samples, bias ratios, or description of each dataset).

Baseline. We compare the performance of our A^2 with state-of-the-art debiasing methods, including empirical risk minimization (ERM). The unsupervised debiasing and augmentation techniques are taken into consideration when choosing the baselines. Detailed descriptions of each baseline method are provided below:

1) ERM. Empirical risk minimization (ERM) indicates a classifier trained with only original CE loss [16], not exploiting any debiasing scheme. ERM performance is treated as a reference to other debiasing methods.

2) LfF. LfF is a state-of-the-art method proposed by Nam et al. [23], utilizing weighted cross-entropy loss for bias-conflict samples in an end-to-end manner. Specifically, LfF trains two networks, a biased classifier for calculating relative difficulty and a debiased classifier for learning debiased representation. After training, the only debiased classifier is used to predict.

3) DisEnt. DisEnt proposed by Lee et al. [19] is the first to apply augmentation approach to the debiasing method; specifically, DisEnt is an extended version of LfF in terms of augmentation. They follow the training mechanism of LfF and introduce the additional swapping function of the feature vectors from the biased classifier.

Implementation Details. We adopt common implementations for the debiasing process: we use a 3-layer MLP network with 100 hidden units for the CMNIST dataset, and Resnet18 [11] network for the other datasets. For training the biased classifier, we set the controlling degree q as 0.7 in equation 1. We set $k = 10$ for our entire experiments for extracting bias-conflict except one-shot testing scenario. See Appendix A.3 for more details.

6 Results

6.1 Performance on Benchmark Datasets

Real-World Datasets. The results are presented in Table 2, where we provide both bias-align and bias-conflict performances.

For the BFFHQ dataset, the ERM model correctly identified bias-align samples, but incorrectly predicted bias-conflict samples, with 39.87% and 50.80% in the one-shot and 99.5% bias setting, respectively. On the other hand, our classifier, trained with the augmented dataset, outperforms ERM and other state-of-the-art debiasing methods by a large margin, producing 47.87% and 56.73% accuracy in the one-shot and 99.5% bias settings, respectively. We can clearly observe that our method is much more effective when the bias setting is severe (one-shot and 99.5%). Furthermore, our A^2 outperforms other baselines in the

Table 2: Performance comparison on both real-world and synthetic datasets. We evaluate both bias-align and bias-conflict test sets to show that each method can learn debiased representation without performance degradation for bias-align samples. We report the average accuracy over three runs. Bold and underlined numbers indicate the best performance and the second best performance, respectively.

Dataset Type	Dataset	Bias Ratio	Bias-align				Bias-conflict			
			ERM	LfF	DisEnt	(Ours)	ERM	LfF	DisEnt	(Ours)
Real-world	Biased FFHQ	One-Shot	99.33	98.67	**99.33**	87.73	39.87	39.47	40.40	**47.87**
		99.5%	99.40	96.67	97.13	**98.93**	50.80	55.73	52.67	**56.73**
	BAR	100%	N/A				63.10	67.84	66.46	**71.15**
Synthetic	Colored MNIST	One-Shot	98.40	**97.80**	96.50	97.21	14.22	22.06	18.33	**23.70**
		99.5%	99.90	83.70	68.33	**93.03**	33.36	53.43	60.96	**67.47**
		99%	99.90	86.65	78.35	**97.14**	57.28	61.75	**75.99**	70.68
		98%	99.83	90.21	90.64	**99.57**	73.24	68.65	**79.69**	76.93
		95%	99.67	85.99	98.21	**99.20**	83.88	85.13	**87.29**	86.09
	Corrupted CIFAR10	One-Shot	97.60	94.97	**96.93**	82.40	9.03	9.44	13.39	**15.45**
		99.5%	96.87	82.47	89.37	**90.03**	13.23	14.43	14.58	**15.96**
		99%	97.27	85.67	**95.63**	82.37	13.46	19.01	19.97	**21.45**
		98%	96.57	78.93	**95.17**	88.53	17.62	**26.50**	23.11	21.79
		95%	95.10	72.93	92.37	**93.70**	27.37	**35.20**	30.39	28.23

most severe case, where the bias ratio is set to 100% in the BAR dataset. We believe this is because our method easily adapts bias-conflict features in the severe bias setting. Overall, our augmentation process demonstrated its effectiveness in a severely biased environment by converting bias-align samples into bias-conflict samples, effectively assisting the model to learn the debiased representations in real-world datasets.

Synthetic Datasets. To further demonstrate the effectiveness of augmentation methods, we evaluate the classification performance in controlled environments with synthetic datasets. For the CMNIST dataset, our classifier outperformed two baselines (ERM and LfF) with bias-conflict samples across all bias ratios. Furthermore, our classifier successfully predicted bias-align samples even after learning the debiased representation, producing on par performance compared to the ERM model. DisEnt performed better than our method in three cases (99%, 98%, and 95%); however, our classifier still outperforms DisEnt with a large margin in extremely biased environments (one-shot and 99.5%). For the CCIFAR10 dataset, our classifier also achieved the state-of-the-art performance in severely biased settings (one-shot, 99.5%, and 99%), and produced on par performance compared to baselines in other cases. In some cases, DisEnt has better accuracy on bias-align samples. We believe that the cases where DisEnt performed better are due to the difficulty of learning debiased representations of bias-conflict samples. In fact, the performance difference between the bias-align and bias-conflict test set in these cases is lower than other baselines across all datasets. Overall, our approach performed better than other approaches across different datasets with various bias ratios, achieving the best performance among one-shot and 99.5% settings and the second best performance among most cases.

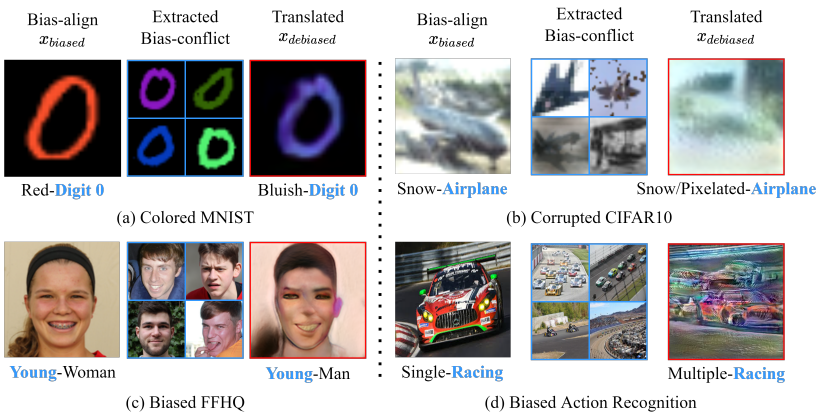

Fig. 4: Qualitative evaluation for the translated images via A^2 with 99.5% bias ratio. The 2^{nd} column (middle images) from each dataset indicates the samples extracted by our extraction algorithm. The blue text indicates label information for each image. We observe that the translated images successfully include bias-conflict features while retaining their respective labels.

6.2 Qualitative Analysis on Augmented Samples

We examine translated images used in our main experiments with 99.5% of bias-align samples. Figure 4 shows results on the bias-align sample for each dataset and the corresponding translated image using the extracted samples. We observed that the translated images have bias-conflict representations against bias-align samples. For example, red-colored digit 0 in CMNIST is translated to a bluish color digit 0.

Interestingly, the translated image has a mixture of several colors from the extracted bias-conflict samples, which implies our A^2 reflects the features from bias-conflict samples. We also observed this successful translation in BFFHQ. Nevertheless, a failure case was also observed in CCIFAR10 when enforcing all noise biases from the conflicting samples to the given image, challenging to recognize the original object. For the BAR dataset, where training samples are all aligned with a bias, we observed that there is an unintended bias in the aligned samples (e.g., most of the images have single racing cars, but extracted images have several cars). Therefore, our A^2 converted a single car image to several car images with the original red-colored car preserved. It is worthwhile to note that our method can also relieve the unintended biases when there are no bias-conflict samples.

6.3 Ablation Study

We conduct ablation studies with our A^2 in terms of the extraction method and the quality of the translated samples, the results are presented in Table 3. We

Bias Ratio	CMNIST	CCIFAR10	BFFHQ
99.5	100	80	60
99.0	100	100	N/A
98.0	100	100	N/A
95.0	91	100	N/A

(a) Bias-Conflict Ratio (%)

Bias Ratio	CMNIST	CCIFAR10	BFFHQ
99.5	98.37	97.03	98.52
99.0	98.95	96.11	N/A
98.0	98.61	96.67	N/A
95.0	98.66	97.95	N/A

(b) High-Quality Ratio (%)

Table 3: A^2 ablation experiments with benchmark datasets. We report (a) the bias-conflict ratio extracted by our extraction method, and (b) the high quality ratio of the translated samples based on the GIQA [9]. We utilize all possible bias ratios for CMNIST, CCIFAR10 and BFFHQ datasets. We cannot include BAR dataset, as it does not have bias-conflict samples in training set.

report all valid cases for CMNIST, CCIFAR10, and BFFHQ datasets except BAR dataset, as BAR does not have bias-conflict samples in the training data.

Bias-conflict Ratio of Extracted Samples. To verify the effectiveness of extracting bias-conflict samples, we measure the proportion of bias-conflict samples among the extracted samples across all datasets. The results are reported in Table 3a. The bias-conflict ratio is surprisingly high, showing mostly from 80% to 100% for CMNIST and CCIFAR10 datasets; however, the ratio was relatively low in the BFFHQ dataset.

We noticed that extracting whole bias-conflict samples in real-world datasets with our proposed method remains challenging. It is difficult because of the highly entangled representations, where correlated complex attributes, such as age and gender, cannot be fully disentangled by the proposed biased classifier. However, we believe that our unsupervised extraction method offers a simple yet practical approach for extracting bias-conflict samples, by efficiently extracting bias-conflict samples without any supervision. Moreover, we successfully enhanced the classification performance in real-world datasets, achieving state-of-the-art results.

High-Quality Ratio of Translated Samples. We measure the percentage of high-quality samples in the translated samples for all datasets in order to assess the quality of the translated bias-conflict samples generated by our A^2. We use generated image quality assessment (GIQA) proposed by Gu et al. [9] for this analysis. Specifically, we use the GMM-GIQA among the GIQA family (refer to the Section 3.2 for more details [9]).

The GMM-GIQA value is in the range of [0, 1], and higher value indicates higher quality. We define a generated image as *high-quality* when it gets the GMM-GIQA value over 0.5. As reported in Table 3b, the majority of samples that generated by our method are evaluated as *high-quality* samples in all bias ratios, producing over 96% results.

We noticed that our framework could enhance the classification performance, providing lower bounds and valid results without using the quality assessment method. Furthermore, our framework can be improved by reducing errors with

the GIQA method (e.g., the GIQA method can be used as an 'image picker,' selecting only high-quality images).

7 Discussion

Limitations. We observed possible limitations of our model during experiments. Since our method requires a biased generative model, dependence on the generative model may arise. However, developing generative models is one of the most fast-growing research fields in machine learning; there is a great possibility for future improvement when combined with our novel approach.
Future Work & Broader impacts. We plan to extend our framework to prevent and filter out badly generated samples by using GIQA metric as the 'image picker' to select the high-quality samples. As machine learning models become deeply embedded in diverse aspects of our daily lives, it is crucial to ensure they produce accurate, reliable, and trustworthy results. In this context, we believe that our approach can contribute to various computer vision applications, such as media or database platforms, by debiasing the massive data for protecting machine learning models from bias.

8 Conclusion

In this work, we proposed a novel augmentation framework, A^2, which effectively learns biased representations through image-level augmentation to address dataset biases. Our framework is derived from our findings that augmenting bias-conflict samples is crucial in biased contexts. Thus, A^2effectively augments bias-conflict samples through image-to-image translation methods integrated with an unsupervised extraction algorithm. We demonstrated the performance and effectiveness of our augmentation framework through extensive experiments. We believe our work can contribute to building more accurate and trustworthy computer vision applications by effectively preventing bias, predominately occurring in real-world datasets.

Acknowledgements We thank members of DASH Lab. for the helpful feedback. This work was partially supported by the Basic Science Research Program through National Research Foundation of Korea (NRF) grant funded by the Korean Ministry of Science and ICT (MSIT) under No. 2020R1C1C1006004 and Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean MSIT: (No. 2022-0-01199, Graduate School of Convergence Security at Sungkyunkwan University), (No. 2022-0-01045, Self-directed Multi-Modal Intelligence for solving unknown, open domain problems), (No. 2022-0-00688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes), (No. 2021-0-02068, Artificial Intelligence Innovation Hub), (No. 2019-0-00421, AI Graduate School Support Program at Sungkyunkwan University), and (No. 2021-0-02309, Object Detection Research under Low Quality Video Condition).

References

1. Agarwal, V., Shetty, R., Fritz, M.: Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9690–9698 (2020)
2. Bahng, H., Chun, S., Yun, S., Choo, J., Oh, S.J.: Learning de-biased representations with biased representations. In: International Conference on Machine Learning. pp. 528–539. PMLR (2020)
3. Cadene, R., Dancette, C., Cord, M., Parikh, D., et al.: Rubi: Reducing unimodal biases for visual question answering. Advances in neural information processing systems **32** (2019)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Clark, C., Yatskar, M., Zettlemoyer, L.: Don't take the easy way out: ensemble based methods for avoiding known dataset biases. arXiv preprint arXiv:1909.03683 (2019)
6. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
7. Goel, K., Gu, A., Li, Y., Ré, C.: Model patching: Closing the subgroup performance gap with data augmentation. arXiv preprint arXiv:2008.06775 (2020)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
9. Gu, S., Bao, J., Chen, D., Wen, F.: Giqa: Generated image quality assessment. In: European conference on computer vision. pp. 369–385. Springer (2020)
10. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
13. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
15. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
16. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: 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)
17. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

18. LeCun, Y., Cortes, C.: MNIST handwritten digit database (2010), http://yann.lecun.com/exdb/mnist/
19. Lee, J., Kim, E., Lee, J., Lee, J., Choo, J.: Learning debiased representation via disentangled feature augmentation. Advances in Neural Information Processing Systems **34** (2021)
20. Li, Y., Vasconcelos, N.: Repair: Removing representation bias by dataset resampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9572–9581 (2019)
21. Lu, D., Weng, Q.: A survey of image classification methods and techniques for improving classification performance. International journal of Remote sensing **28**(5), 823–870 (2007)
22. Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence (2021)
23. Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J.: Learning from failure: De-biasing classifier from biased classifier. Advances in Neural Information Processing Systems **33**, 20673–20684 (2020)
24. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10743–10752 (2021)
25. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints pp. arXiv–1807 (2018)
26. Puli, A.M., Zhang, L.H., Oermann, E.K., Ranganath, R.: Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In: International Conference on Learning Representations (2021)
27. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
28. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019)
29. Van Dyk, D.A., Meng, X.L.: The art of data augmentation. Journal of Computational and Graphical Statistics **10**(1), 1–50 (2001)
30. Wang, H., He, Z., Lipton, Z.C., Xing, E.P.: Learning robust representations by projecting superficial statistics out. arXiv preprint arXiv:1903.06256 (2019)
31. Ying, X.: An overview of overfitting and its solutions. In: Journal of Physics: Conference Series. vol. 1168, p. 022022. IOP Publishing (2019)
32. Zhou, C., Ma, X., Michel, P., Neubig, G.: Examining and combating spurious features under distribution shift. In: International Conference on Machine Learning. pp. 12857–12867. PMLR (2021)
33. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=6xHJ37MVxxp
34. Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055 (2019)