From Fake to Real: Pretraining on Balanced Synthetic Images to Prevent Spurious Correlations in Image Recognition

Maan Qraitem[1], Kate Saenko[1], and Bryan A. Plummer[1]

[1]Boston University
{mqraitem,saenko,bplum}@bu.edu

Abstract. Visual recognition models are prone to learning spurious correlations induced by a biased training set where certain conditions B (*e.g.*, Indoors) are over-represented in certain classes Y (*e.g.*, Big Dogs). Synthetic data from off-the-shelf large-scale generative models offers a promising direction to mitigate this issue by augmenting underrepresented subgroups in the real dataset. However, by using a mixed distribution of real and synthetic data, we introduce another source of bias due to distributional differences between synthetic and real data (*e.g.* synthetic artifacts). As we will show, prior work's approach for using synthetic data to resolve the model's bias toward B do not correct the model's bias toward the pair (B, G), where G denotes whether the sample is real or synthetic. Thus, the model could simply learn signals based on the pair (B, G) (*e.g.*, Synthetic Indoors) to make predictions about Y (*e.g.*, Big Dogs). To address this issue, we propose a simple, easy-to-implement, two-step training pipeline that we call From Fake to Real (FFR). The first step of FFR pre-trains a model on balanced synthetic data to learn robust representations across subgroups. In the second step, FFR fine-tunes the model on real data using ERM or common loss-based bias mitigation methods. By training on real and synthetic data separately, FFR does not expose the model to the statistical differences between real and synthetic data and thus avoids the issue of bias toward the pair (B, G). Our experiments show that FFR improves worst group accuracy over the state-of-the-art by up to 20% over three datasets. Code available: https://github.com/mqraitem/From-Fake-to-Real

Keywords: Spurious Correlations · Synthetic Data Augmentation

1 Introduction

Visual recognition models are prone to learning spurious correlations (Bias) [22,35,40]. These correlations frequently arise due to an imbalance in the training set. For example, given a dataset with classes Y (*e.g.*, Smiling vs Not Smiling), there exists a confounding bias variable B (Gender: Male and Female) in the training set such that one bias group (*e.g.*, Male) is represented in one class more than others (*e.g.*, most males are Smiling). This leads models to mistakenly use the bias signal B (gender) to predict Y (Smiling). Rapid progress in

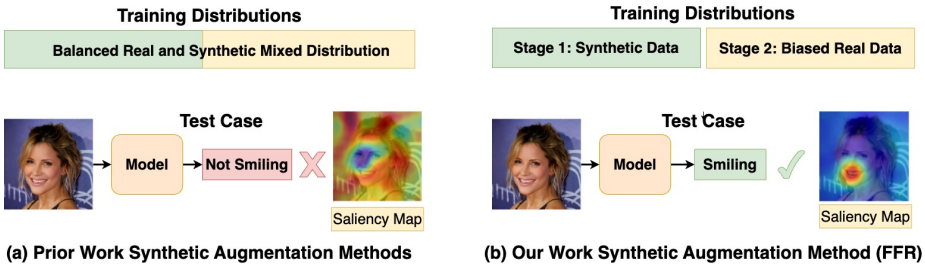


Fig. 1: Comparison of RISE [24] saliency maps produced for a model trained to predict the attribute Smiling Y given a bias toward Gender B (Most Women are not Smiling) using: (a) prior work in Synthetic Augmentation [23, 26], which do not address the unexpected bias toward the pair (Gender B, Data Source G), *e.g.*, (Female-Synthetic vs Female-Real) and, thus, use spurious features leading to an incorrect prediction (not smiling) (b) our approach FFR where the Synthetic and Real Data are separated into two training stages, thereby mitigating the bias suffered by prior work enabling us to learn the correct features (Mouth) and correctly predict Smiling.

large-scale generative models, notably diffusion-based models [7, 31], provides a clear mitigation method that alleviates bias using synthetic data: use a mixed distribution of synthetic and real data that alleviates the real dataset bias.

Prior work has introduced several methods to achieve this goal. For example, Additive Synthetic balancing (ASB) [26, 33] augments the biased real dataset with a balanced synthetic dataset. Uniform Synthetic Balancing (USB) generates enough data to uniformly balance the dataset subgroups [23, 36], *i.e.*, each subgroup will have the same number of samples[1]. However, by training the real and synthetic data samples at the same time, a model may simply learn to identify correlations between bias B and whether the data was real or generated G (*e.g.* by using generative model artifacts [3]). For example, in the setting where the training data contained mostly smiling men but few smiling women, prior work may simply learn that synthetically generated women smile (but women in real images do not). Thus, as shown in Figure 1(a), models trained using strategies of prior work (*e.g.*, ASB and USB) may focus on unrelated features for the target task. In addition, assuming some distributional differences between synthetic and real data, we provide a theoretical analysis that shows that every possible augmentation of a biased dataset with synthetic data will exhibit some bias toward (B, G); *i.e.*, $P_D(Y|B, G) \neq P_D(Y)$.

To mitigate this problem, we rethink how synthetic data is used for bias mitigation by developing a simple, easy-to-implement, yet effective two-stage training pipeline called From Fake to Real (FFR). The first stage involves pretraining on balanced synthetic data where we learn robust representations across subgroups. In the second step, FFR fine-tunes the model on real data using ERM or common loss based bias mitigation methods [8, 10, 28, 29, 34]. By separating

[1] Refer to the supplementary for a visual comparison of prior work in synthetic augmentation (*e.g.*, [23, 26, 33, 36]) and our approach.

the two data sources (*i.e.* Real and Synthetic) into two different training stages, FFR doesn't expose the model to the statistical differences between real and synthetic data (*e.g.* generative model artifacts [3]) and, thus, avoids the issue of bias that might arise from training on these two sources of data together. Effectively, the synthetic data acts as a source of unbiased representations for each subgroup, leading to improved performance when training with the real data using ERM or loss-based bias mitigation methods in the second step. As shown in Figure 1(b), this enables FFR to learn more relevant features rather than focusing on spurious background features.

To evaluate our approach, we expand on the experimental frameworks of prior work, which are limited to one bias rate per dataset [9, 25, 30]. Instead, we conduct systemic analysis over three datasets, CelebA-HQ [14], UTK-Face [39], and SpuCo Animals [9], and a range of bias rates ranging from moderate to severe resulting in over 5k experiments in total.

Our contributions are summarized below:

- We introduce a simple, easy to implement, yet effective, two-step training pipeline (FFR) that uses synthetic data to alleviate the issue of spurious correlations (Bias). Unlike prior work, our pipeline avoids the issue of bias to distributional differences between real-synthetic data (*e.g.*, generative model artifacts) and, thus, is more effective at mitigating bias.
- We provide a theoretical analysis on how augmentation with synthetic data results in an unexpected bias toward synthetic artifacts.
- Comprehensive experiments over three datasets (UTK-face, CelebA-HQ, and SpuCo Animals) and at least four bias strengths per dataset validate our method's effectiveness. Indeed, FFR improves performance over state-of-the-art worst accuracy by up to 20%.

2 Related Work

Mitigating Bias with Synthetic Data. As noted in the Introduction, some limited work exists on using synthetic data augmentation to address issues due to imbalanced training data. This includes Uniform Synthetic Balancing (USB) [23, 36], which balances underrepresented subgroups, where subgroups are the intersection of classes Y and bias groups B. This, in turn, effectively ensures that Y is statistically independent from B, *i.e.*, $P_{\bar{D}}(Y|B) = P_{\bar{D}}(Y)$ where $\bar{D}$ is the combined dataset of real and synthetic data. Additive Synthetic Balancing (ASB) [26,33] augments a biased real dataset with a balanced synthetic dataset. In our work, we show how both approaches (USB and ASB) result in models that are biased toward (B, G) where $G = \{Real, Synthetic\}$, *i.e.*, the variable that differentiates between real and synthetic data. We could attempt to mitigate this issue by combining USB and ASB with loss-based bias mitigation methods (*e.g.*, [8, 10, 28, 29, 34]). However, in order to account for the new source of bias from (B, G) where $G = \{Real, Synthetic\}$, this approach doubles the number of bias groups ($|(B, G)| = |B||G| = 2|B|$) which increases the optimization difficulty, reducing performance as we will show in Section 4.1. Instead, our two-stage

training pipeline addresses the issue of new biases being introduced from using synthetic data by training both real and generated data separately.

Synthetic-Data-Free Mitigation Methods. Also related to our task are methods that use architecture changes and/or alter the training procedures to mitigate dataset bias [8, 10, 28, 30, 34, 37]. For example, Sagawa *et al.* [30] proposes GroupDRO (Distributionally Robust Neural Networks for Group Shifts), a regularization procedure that adapts the model optimization according to the worst-performing group. More recently, a series of works seeks to mitigate bias assuming no access to bias labels in training time [1, 12, 18, 38]. Most recently, DFR [12] showed how fine-tuning a model on a small balanced validation (after being trained on the biased training set) achieves state-of-the-art performance. Our work complements these efforts by introducing a novel pipeline for using synthetic data that further boosts the performance of these methods, especially in high-bias settings. Thus, as we will show, these methods and our approach can be combined to boost performance over either when they are used alone.

Uncovering Spurious Correlations. In our work, we are interested in mitigating spurious correlations; a spurious correlation results from underrepresenting a certain group of samples (*e.g.*, samples with the color red) within a certain class (*e.g.*, planes) in the training set. This leads the model to incorrectly correlate the class with the over-represented group. For example, prior work has shown that several datasets exhibit spurious correlations [6, 15, 22]. For example, Meister *et al.* [22] reports that models trained on COCO [16] and OpenImages [13] learn spurious correlations with respect to various gender artifacts. Li *et al.* [15] showed that models trained on ImageNet spuriously correlate the Carton class with Chinese watermarks. Hirota *et al.* [6] showed how several VQA dataset encodes racial and gender biases. Our work complements these effort by introducing a more effective way of using synthetic data to mitigate spurious correlations.

3 Synthetic Data for Robust Representations against Bias in Image Recognition

Visual classification models can often rely on spurious correlations in the training set that do not reflect their real-world distribution. More concretely, given a dataset of images X, classes Y, and bias signal B (*e.g.*, Gender: Male/Female), a biased model relies on the signal in X that infer B to make predictions $\hat{Y}$. This is because the distribution $P_D(Y|B) \neq P_D(Y)$, *i.e.*, the training set encodes some correlation between the classes and the biases. For example, given a class y (*e.g.*, Smiling), a certain bias group b (*e.g.*, Male) might be over-represented compared to others. Therefore, a model might mistakenly predict the class of an image (*e.g.*, Not Smiling) as the wrong class (*e.g.*, Smiling) because the signal b (Male) is present in the image (*e.g.*, Man is Not Smiling) [25, 29].

To address this issue, our work explores using synthetic data from generative models as we will discuss in detail below. Section 3.1 explores how augmenting the real dataset with synthetic data results in a bias towards distributional

differences between synthetic and real data. Section 3.2 introduces From Fake to Real (FFR); our novel two-stage pipeline that addresses this issue.

3.1 Motivation

In this section, we explore a critical problem with the class of solutions that mitigates dataset bias by augmenting biased datasets with synthetic data, *e.g.*, Additive Synthetic Balancing (ASB) [26, 33] and Uniform Synthetic Balancing (USB) [23, 36]. These approaches don't consider the fact that the distribution of synthetic data is not the same as the distribution of real data. Indeed, while research on generative models has made significant progress in producing ever more realistic images, especially with the recent advent of diffusion models [7,31], there might still be some distributional differences between the real and synthetic data. For example, Corvi *et al.* [3] demonstrates how state-of-the-art diffusion models leave fingerprints in the generated images that recognition models could use to differentiate between real and synthetic data.

Assuming real and synthetic data are drawn from different distributions, and we are given a biased dataset D, *i.e.* $P_D(Y|B) \neq P_D(Y)$, we argue that it is impossible to guarantee that we can create $\bar{D}$ where Y is not biased toward the pair (B, G). Formally:

Theorem 1. *Assume we are given dataset D where $P_D(Y|B) \neq P_D(Y)$ such that Y are target labels and B are biased group labels (*i.e. *dataset is biased). Assume $\bar{\mathcal{D}}$ represent all possible versions of the dataset augmented with synthetic data such that $G = \{Real, Synthetic\}$, then for every $\bar{D} \in \bar{\mathcal{D}}$, $P_{\bar{D}}(Y|B, G) \neq P_{\bar{D}}(Y)$ where G are the synthetic/real labels.*

Refer to the supplementary for a proof. As shown, this Theorem guarantees that it is impossible to create an augmented version of the dataset D, *i.e.*, $\bar{D}$, without $\bar{D}$ exhibiting some bias toward (B, G). Therefore, this implies that both methods from prior work, ASB [26, 33] and USB [23, 36], may rely on biased signals stemming from (B, G) to make predictions.

To gain some intuition, consider the following illustrative example for Uniform Synthetic Balancing (USB): in an attempt to mitigate the dataset bias of class Landbirds being mostly on Land and Waterbirds being most Water, a significant number of synthetic samples of Landbirds on Water and Waterbirds on Land are added to the dataset. While this means that there is an equal number of Landbirds and Waterbirds on Land and on Water in the combined dataset, *i.e.*, $P_{\bar{D}}(Y|B) = P_{\bar{D}}(Y)$, this also means that there are significantly more *Synthetic* Landbirds on Water than there are *Synthetic* Landbirds on Land. Assuming that the model could differentiate between real and synthetic images, then it is likely advantageous to learn the signal pair (Water, Synthetic) in order to predict the class Landbird while the signal (Water, Real) predicts the class Waterbirds.

3.2 From Fake to Real (FFR): A Two-Stage Training Pipeline

Our approach, From Fake to Real (FFR), aims to address the issue in prior work outlined in Section 3.1, where models learn a bias between the target labels Y

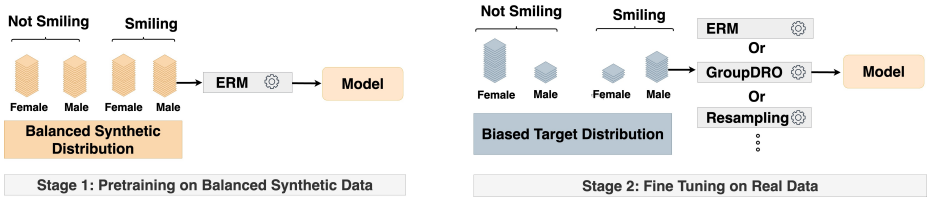


Fig. 2: An overview of From Fake to Real (FFR) that incorporates synthetic data to mitigate bias. In Stage 1, we pretrain on a balanced synthetic dataset where we learn robust representations across subgroups. In Stage 2, we fine-tune the model on real data using ERM or common synthetic-data-free bias mitigation methods. By training on real and synthetic data separately, FFR does not expose the model to the statistical differences between real and synthetic data and thus avoids the issue of bias between the two data sources. Refer to Section 3.2 for further discussion.

and the pair labels (B, G). The key to our approach is the separation of training on the two data sources, real and synthetic, into two different stages. The model is exposed to one data source at a time, which effectively prevents the use of signals from the pair (B, G) to make predictions as neither appear in the same training step. We provide additional details on our two training stages below:

Step 1: FFR pretrains a model M on a balanced synthetic dataset D_{syn} where $P_{D_{syn}}(Y|B) = P_{D_{syn}}(Y)$. To obtain this distribution, we simply deploy a generative model to sample the same number of synthetic data per bias subgroup. This step enables the model M to learn robust initial representations for each subgroup. Refer to Figure 2 (Stage 1) for an overview of this step. Denote the resulting model from this step as $\bar{M}$.

Step 2: While Step 1 learns valuable unbiased representations, there is still a distribution shift going from real to synthetic datasets [32]. Therefore, we fine-tune the model $\bar{M}$ from Step 1 on the real dataset to better fit to its distribution. We find that even a simple empirical-risk minimization fine-tuning using the model $\bar{M}$ as an initialization is sufficient to boost performance. However, the real dataset's distribution D is biased, *i.e.*, $P_D(Y|B) = P_D(Y)$. Thus, some of the benefits of our first stage pretraining are undone as the model might simply relearn the bias. To address this, we combine our two-stage training pipeline with loss-based bias mitigation methods (*e.g.*, [8, 10, 28, 29, 34]). Refer to Figure 2 (Stage 2) for an overview of this step. As we note in our experiments, regardless of the method used in Step 2, we observe a significant performance boost using Step 1's model $\bar{M}$ for initialization.

In summary, FFR is a flexible framework that rethinks the use of synthetic data for bias mitigation. We use FFR to deploy synthetic data to learn initial unbiased representations to improve the performance of training on real data regardless of the method used to train on real data. Therefore, it is generalizable to any bias mitigation method and easy to implement no matter the model architecture. Finally, our framework effectively avoids the issue of bias to distributional differences between real and synthetic data, unlike prior work's methods.

4 Experiments

Datasets. We seek to compare the effect of synthetic augmentation on varying amounts of bias. To that end, we use three standard bias mitigation datasets, namely: CelebA-HQ [14], SpuCo Animals dataset [9], and UTK-Face dataset [39]. SpuCo Animals has one possible bias variable ("Background") where the bias is 95% (*i.e.*, the majority bias group takes up 95% of the class distribution). Prior work that used CelebA [30] used the attribute "Blonde Hair," which has $\sim 97\%$ bias, and "Wearing Lipstick," which has $\sim 99.9\%$ bias. In contrast, prior work that used Utk-Face [25] used the Age attribute with 90% bias. However, since we seek to study the effect of different bias settings on the methods' performance, simply comparing performance on different attributes with different biases is not fair, as it entangles the difficulty of learning different targets (*e.g.*, "Wearing Lipstick" vs. "Blonde Hair") and the difficulty of learning different bias ratios (97% vs. 99.9%). To mitigate this issue, we choose to fix the bias and target attribute per dataset and manually vary the bias according to 5 main ratios ranging from moderate to severe $(90\%, 95\%, 97\%, 99\%, 99.9\%)$ by simply dropping samples appropriately from the minority groups to match the target bias ratio. Specifically, we evaluate using: 1) CelebA-HQ [14], where we choose "Smiling" as the target attribute and "Gender" as the bias attribute, 2) UTK-Face [39], where we use "Age" as the bias attribute and "Gender" as the target attribute, and 3) SpuCo Animals [9], where the bias attributes are {*Indoors, Outdoors, Land, Water*} and target attributes are {*Small dogs, Big Dogs, Landbirds, Waterbirds* }. Note that SpuCo Animals has a minimum bias of 95%. Thus, 90% bias is only used for UTK-Face and CelebA-HQ.

Metrics. Following [30], we use Worst Accuracy (WA) to measure the models' spurious behavior. This metric returns the accuracy of the worst-performing subgroup where the subgroup is defined as the intersection of class and bias groups. In addition, we use balanced accuracy (BA), which averages the accuracies of all subgroups [25]. BA reflects the overall performance of the model while not being biased by the majority of subgroups.

Implementation Details. We use a Resnet50 [5] backbone trained with ADAM [11], where we use grid search to set the learning rate over the validation set. We use default values for the other parameters. Furthermore, we find that freezing the batch norm in FFR Stage 2 to be helpful on some datasets. See the supplementary for additional implementation details for each method. For data generation, we use Stable Diffusion V1.4 [27], where we use the prompt template *A photo of* ***{bias}*** ***{class}*** to sample new images. As a powerful generator is not easily accessible for every application, in the supplementary we explore the effect of the quality of the synthetic images on performance.

Methods. We report the performance of training with three modes of incorporating synthetic data:

- **None**: No synthetic data is used.
- **USB** [23]: Synthetic data is used to uniformly balance the distribution
- **ASB** [26]: A balanced synthetic dataset is added to the real dataset (ASB)

– **FFR**: our method where we first pre-train on balanced synthetic data using ERM and then fine-tune on real data.

For each mode, we report performance using Empirical Risk Minimization (ERM) and several popular state-of-the-art bias mitigation methods. More concretely, we report the performance of GroupDRO [30], Resampling, and Deep Feature Reweighting (DFR) [12]. GroupDRO is an optimization technique where the contribution of each subgroup loss is weighted by their performance. Resampling oversamples minority subgroups such that each subgroup is equally represented per batch. DFR fine-tunes a linear layer over the feature space on a balanced validation set. Note that Group-DRO and Resampling require access to Bias labels in the training set, while DFR does not.

Note that when no synthetic data is used, the bias mitigation methods are deployed to minimize the bias toward B. When combined with USB and ASB, the methods are deployed to mitigate the bias toward (B, G). Finally, when the methods are combined with FFR, they are deployed to mitigate bias toward B only since FFR minimizes the bias toward G by definition. Finally, we fix the number of synthetic data samples used for each method: USB, ASB, and FFR. The fixed size is the number of samples required to balance the dataset in USB.

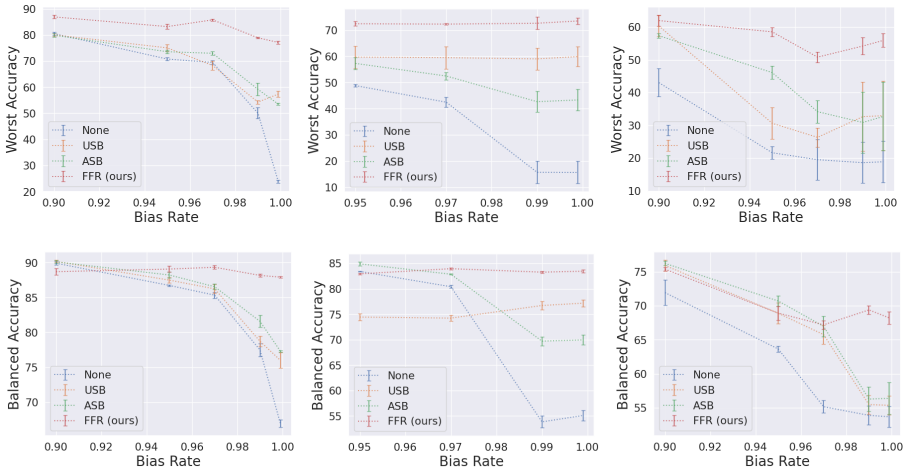
4.1 Comparing Synthetic Augmentation Methods with ERM

Figure 3 compares the performance of our Synthetic Data Augmentation method (FFR) to prior work methods (USB [23] and ASB [26]) over three datasets and various bias ratios. Note how our method (FFR) either matches or improves the worst and balanced accuracy of ASB and USB over each dataset and each bias ratio. For example, FFR improves over USB and ASB on UTK-Face and Bias ratio 95% by over 10%. This is because, as we discuss in Section 3.2, FFR addresses the bias between real and synthetic data and, thus, is more able to use both data sources to mitigate the bias effectively.

More notably, we find that the augmentation methods of prior work result in stable performance on SpuCo Animals across bias ratios, but their performance decreases significantly as bias increases on CelebA-HQ and UTK-Face. This is unlike our method, where the performance remains stable. This demonstrates that our method is more robust to more severe bias.

4.2 Combining Synthetic Augmentation Methods with Synthetic-Data-Free Bias Mitigation Methods.

In this Section, we combine the synthetic data augmentation methods, namely prior work USB [23] and ASB [26], and our method FFR with synthetic-data-Free bias mitigation methods: GroupDRO [30], DFR [12], and Resampling. As we noted at the beginning of Section 4 under Baselines, when combining these methods with ASB and USB, we deploy the synthetic-data-free bias mitigation methods to address the bias toward both (B, G). However, when deployed with FFR, they are implemented to address the bias toward B (the bias toward G is

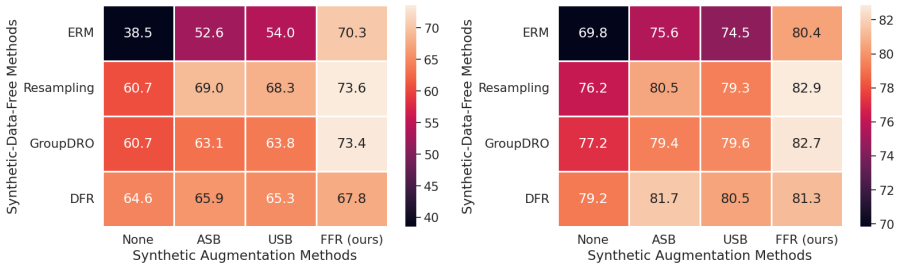


(a) CelebA-HQ [14] **(b)** SpuCO Animals [9] **(c)** UTK-Face [39]

Fig. 3: Comparison of performance between the effect: (None) no synthetic data is used, (USB) synthetic data is used to uniformly balance the distribution (extension of prior work on imbalanced classification [23]), (ASB) balanced synthetic data is added to the real dataset [26] and (FFR) our method where pretrain on balanced synthetic data and fine tune on real data. Models are trained with ERM. Note how our method either matches or improves the performance of prior work augmentation methods. Refer to Section 4.1 for discussion.

automatically addressed by FFR (Section 3.2)). Figure 4 reports the averaged performance over the three datasets and all bias ratios. For simplicity, denote the synthetic-data-free methods (GroupDRO, Resampling, DFR) as **SD-Free** and synthetic-data-augmentation methods (USB, ASB, and FFR) as **SD-Aug**. Below, we consider the impact of **SD-Free** on **SD-Aug** (SD-Free $\rightarrow$ SD-Aug). Then we consider the impact of **SD-Aug** on **SD-Free** (SD-Aug $\rightarrow$ SD-Free).

SD-Free $\rightarrow$ SD-Aug Note the change in performance from top to bottom in Figure 4. SD-Free methods significantly improve the performance of SD-Aug methods. For example, the average worst accuracy of ASB improves by 16.4% (goes from 52.6 % to 69.0 %).. We note a similar trend with USB. This is likely because SD-Free methods address some of the bias between data distributions (real and synthetic) that USB and ASB fail to address. However, even when combined with SD-Free methods, ASB and USB still lag behind FFR even when no SD-Free methods are used (namely row 1 in the Figure where FFR shows 70.3 worst accuracy). This is likely because FFR addresses the bias toward (B, G) by definition while SD-Free methods have to deal with double the number of bias groups in B to address the bias toward (B, G). As a result, this renders the optimization procedure more difficult especially in high bias settings where few samples of the minority groups are available. Overall, these results indicate that FFR by itself is a simple yet effective method of mitigating bias. Nevertheless,



(a) Worst Accuracy **(b)** Balanced Accuracy

Fig. 4: Comparing the performance of synthetic-data-free bias mitigation methods, namely GroupDRO [30], Resampling, and Deep Feature Reweighting (DFR) [12] with no synthetic augmentation (None) as well as with synthetic augmentation using prior work methods (USB [23] and ASB [26]) and our method FFR. Performance is averaged across three datasets and five bias ratios. Note how our method (FFR) in column four is best at improving the performance of non-synthetic-data augmentation methods. Refer to 4.2 for further discussion.

when FFR is combined with SD-Free, we see some improvements with Group-DRO and Resampling (by about 3 points on worst accuracy). This is likely the result of addressing some of the bias toward (B) that might be learned in Step 2 of FFR, as discussed in Section 3.2.

SD-Aug → SD-Free Note the change in performance from left to right in Figure 4. Overall, the performance of SD-Free methods improve as a result of using SD-Aug methods. This is likely the result of SD-Aug methods improving the representations of minority groups, especially in high-bias settings where few samples of the minority groups are available. More notably, the worst accuracy *most* improves when using our SD-Aug method (FFR), where it improves Resampling, GroupDRO, and DFR by $13\%, 13\%, 4\%$, respectively. This is because, as we discussed in the previous section, our method automatically addresses the synthetic-real bias (*i.e.* bias toward (B, G)).

4.3 FFR Design Ablations

FFR is composed of two stages. Stage 1: Pretraining on balanced synthetic data and Stage 2: Fine tuning on real data. Pretraining is done using ERM, and Fine tuning could be done with ERM or bias mitigation methods like Group-DRO and Resampling, which yield further improvements as discussed in Section 3.2. In this Section, we study the effect of FFR Stage and Pretraining choices.

FFR Stages. Table 1 reports the performance of FFR using Stage 1 only, Stage 2 only, Stage 2 followed by Stage 1, and then Stage 1 followed by Stage 2. Note that Stage 2 (table line 2) by itself yields poor performance. This is expected as Stage 2 amounts to training a model with ERM on the biased data without pretraining on synthetic data. Thus, the model, as expected, learns the bias. Training with Stage 1 by itself (table line 1) while improving performance over

Table 1: Ablation of FFR stages over SpuCO Animals averaged over all bias ratios. Note how the inclusion of both stages in our chosen order ($1 \rightarrow 2$) achieves the best performance, confirming the importance of our method design. Refer to Section 4.3 for further discussion.

	WA	BA
Stage 1	51.2 $\pm$0.3	77.0 $\pm$0.9
Stage 2	30.6 $\pm$5.4	68.1 $\pm$1.2
Stage 2 $\rightarrow$ Stage 1	59.6 $\pm$5.4	82.8 $\pm$0.9
Stage 1 $\rightarrow$ Stage 2: FFR (ours)	**72.6** $\pm$2.3	**83.4** $\pm$0.3

Table 2: Ablation of the pretraining distribution used with FFR over CelebA-HQ averaged over all bias ratios. Note how using a balanced distribution during pretraining is important to achieve good performance. Refer to Section 4.3 for further discussion.

	WA	BA
No Synthetic Augmentation	58.9 $\pm$1.8	81.2 $\pm$0.9
FFR w/ Biased Pretraining	69.6 $\pm$3.6	84.4 $\pm$1.0
FFR w/ Balanced Pretraining	**82.3** $\pm$1.1	**88.6** $\pm$0.5

Stage 2 by itself doesn't match the performance of FFR (Stage 1 $\rightarrow$ Stage 2). This is likely due to the real and synthetic data distribution gap. Therefore, following up Stage 2 with Stage 1 is important to bridge the performance gap. Finally, note that reversing FFR (table line 3) doesn't match the performance of FFR. This is likely because the model overfits over the synthetic data.

FFR Pretraining. Table 2 compares balanced pretraining to pretraining on a biased synthetic distribution that follows the biased real distribution. Note how the performance drops significantly when a biased distribution is used for pretraining. These results offer compelling evidence that using a balanced synthetic distribution during pretraining is crucial.

4.4 Empirical Investigation of the Real-Synthetic Data bias

The main motivation behind FFR is mitigating the bias that could arise due to distributional differences between Real and Synthetic data (*e.g.*, Synthetic artifacts [3]) when addressing dataset bias. Assuming these differences, we prove in Section 3.1 that every attempt to balance a biased dataset with synthetic data results in a new bias against the pair (B, G) where B denotes the original dataset bias categories and G denotes whether the image is synthetic or real. FFR addresses this issue by simply dedicating a training step for each data source. The positive impact of FFR is evident from the improved worst accuracy performance noted in Section 4.1. In this Section, we seek to further verify this claim through two additional experiments outlined below.

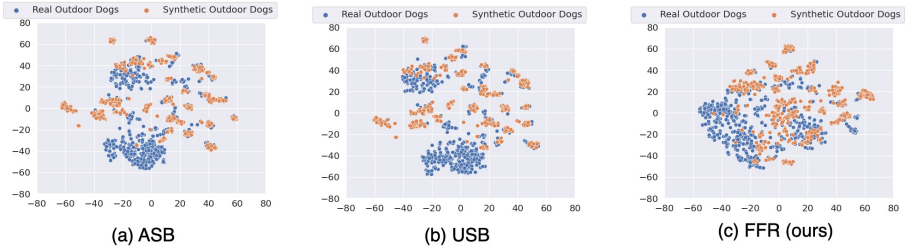


Fig. 5: Comparing the projections of Real vs. synthetic Data using t-SNE [20] with prior work synthetic augmentation (USB [23] and ASB [26]) and our synthetic Augmentation method (FFR). Note how our method (FFR) is the best method for projecting Real and Synthetic data close to each other. This is likely because FFR is less impacted by the bias between real and synthetic data and, thus, is posed to learn best from the two data sources. Refer to Section 4.4 for further discussion.

Table 3: Comparison between Uniform Synthetic Balancing (USB), Additive Synthetic Balancing (ASB), and From Fake to Real (FFR) on Synthetic Data versus Real Data using the SpuCO Animals Dataset. Results are averaged over all bias ratios. Refer to Section 4.4 for discussion.

	Real Data		Synthetic Data	
	WA	BA	WA	BA
USB	59.5 $\pm$8.2	75.6 $\pm$1.3	68.7 $\pm$0.2	82.5 $\pm$0.4
ASB	48.9 $\pm$5.8	76.8 $\pm$1.1	80.7 $\pm$0.3	91.1 $\pm$0.1
FFR (ours)	**72.6** $\pm$2.3	**83.4** $\pm$0.3	**89.2** $\pm$0.2	**96.3** $\pm$0.3

FFR projects Real and Synthetic image embeddings more tightly. If prior work's synthetic augmentation methods are biased with respect to (B, G), then they likely use different features per data source when making a prediction (*e.g.*, the model will use the synthetic artifacts when making predictions about the synthetic data). This, in turn, will likely mean that the synthetic data embeddings are clustered separately from those of real data. However, if the method is not impacted by the real-synthetic bias, then that means that it uses the same correct core features per data source (*e.g.*, features about the dog rather than synthetic artifacts when deciding if the dog is a small dog or a big dog). To verify this claim, in Figure 5 we plot the t-SNE [20] projections of USB [23], ASB [26] and FFR real vs. synthetic embeddings. Note how both ASB and USB clearly project real and synthetic data into two separate clusters. This indicates that ASB and USB likely use features unique to the real vs. synthetic data (*e.g.*, artifacts). However, our method (FFR) projects these samples more tightly indicating that it uses the same unbiased core features when making predictions.

FFR is better at learning from Synthetic Data. If prior work's synthetic augmentation methods are impacted by the bias toward (B, G), then they not only will have learned a biased behavior on the real data, but also on the syn-

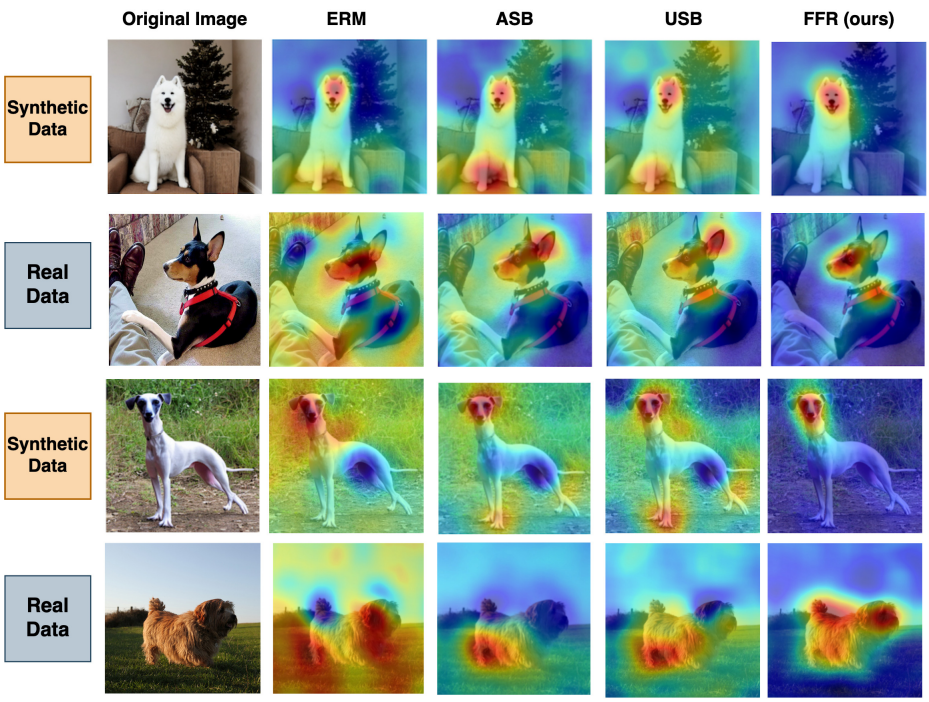


Fig. 6: Saliency maps using RISE [24] when predicting Big Dogs (top two rows) and Small dogs (bottom two rows) using ERM, ASB [26], USB [23, 36] and our method FFR to augment the dataset with synthetic data. The real images are from SpuCO Animals [9], and the synthetic data is from Stable Diffusion v1.4 [27]. Note how our method (FFR) is the only method that can localize the relevant dog features and not get distracted by spurious background features. Refer to Section 4.5 for discussion.

thetic data. To verify this, observe the worst accuracy compared to the balanced accuracy on real versus synthetic data in Table 3. Note how both USB [23] and ASB [26] perform poorly (low worst accuracy compared to balanced accuracy) on both the real and synthetic data. This indicates that rather than using synthetic data to mitigate the bias and generalize to real data, both methods learned to be biased against minority groups on both the real and synthetic data. However, our method (FFR), doesn't suffer from this issue. Specifically, FFR worst-group accuracy on both the synthetic and real data is higher than ASB and USB and closer to FFR balanced accuracy indicating significantly less biased behavior.

4.5 Qualitative Analysis

In this Section, we conduct a qualitative comparison between ERM without any synthetic data, Additive Synthetic Balancing (ASB) [26], Uniform Synthetic Balancing (USB) [23] , and our method From Fake to Real (FFR) on the SpuCo Animals dataset [9] with bias rate 99.9%. Note that the dataset contains four

classes: Big Dogs, Small Dogs, Landbirds, and Waterbirds. In this Section, we focus on the minority subgroups "Big Dogs Indoors" and "Small Dogs outdoors" and sample a real and synthetic image from each subgroup. For each image and model, we produce a saliency map using RISE [24]. Figure 5 reports our results, where we find FFR is the only method that is able to focus on the dog features while disregarding features from the background in both the synthetic and real images. For example, in the second row, both ASB and USB pay attention to the man's feet as well as the ground floor and what seems to be the bottom of a couch to make predictions. Whereas our method (FFR) only focuses on the dog's features. More interestingly, note how for the synthetic images in rows 1 and 3, prior work methods (ASB and USB) use generative artifacts (*e.g.*, three "toes" for the dog rather than four) to make predictions, whereas our method (FFR) ignores these features. Thus, our method is effective at resolving the issue of bias toward the distributional differences between real and synthetic data.

5 Conclusion

We demonstrated through empirical and theoretical work that bias mitigation methods which augment biased datasets with synthetic data fail to address a bias due to distributional difference between real and synthetic data. To address this issue, we introduced From Fake to Real (FFR): a framework that separates training on synthetic data from training on real data, thus, avoiding the bias between the two data sources. Our systemic analysis over three datasets and five bias settings per dataset demonstrated how our method improved worst group accuracy over prior work methods by up to 20%. Furthermore, FFR continued to show superior performance even when methods where combined with synthetic-data-free methods. Finally, we provided an extensive ablation that confirms our methods design choices including the pretraining and stage choices.

Limitations and Future Work In our work, we use large pre-trained text-to-image models to generate synthetic data. While the property of controllable generation using text allows us to generate data that undoes the bias of the real dataset, the generative model might nevertheless inject some biases into the generated data that are not accounted for by the text used to generate the images. For example, Stable Diffusion [27] used in this work has been demonstrated to exhibit several biases [2, 19]. Moreover, as noted in Table 2, FFR relies on the generative model being able to faithfully generate a balanced synthetic dataset to achieve good performance; an imbalanced pretraining distribution significantly hurts performance. However, as prior work noted [17], recent diffusion models occasionally struggle to follow some prompts espeically ones that require compositionality [4, 21]. Therefore, this might jeopardize the ability of diffusion models to generate a balanced pretraining distribution in some cases and thus hurt FFR performance. Therefore, future research that focuses on training fairer and more accurate generative models would alleviate some of these issues. Nevertheless, our approach is generative model agnostic as it is addressing the issue of bias due to data source bias.

Acknowledgments This material is based upon work supported, in part, by DARPA under agreement number HR00112020054 and the National Science Foundation, including under Grant No. 2120322, 2134696. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the supporting agencies.

References

1. Ahmed, F., Bengio, Y., Van Seijen, H., Courville, A.: Systematic generalisation with group invariant predictions. In: International Conference on Learning Representations (2020)
2. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 1493–1504 (2023)
3. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10095167
4. Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., Yang, Y.: Benchmarking spatial relationships in text-to-image generation. arXiv preprint arXiv:2212.10015 (2022)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)
6. Hirota, Y., Nakashima, Y., Garcia, N.: Gender and racial bias in visual question answering datasets. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 1280–1292 (2022)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. ArXiv **abs/2006.11239** (2020)
8. Hong, Y., Yang, E.: Unbiased classification through bias-contrastive and bias-balanced learning. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021), https://openreview.net/forum?id=2OqZZAqxnn
9. Joshi, S., Yang, Y., Xue, Y., Yang, W., Mirzasoleiman, B.: Towards mitigating spurious correlations in the wild: A benchmark & a more realistic dataset. arXiv preprint arXiv:2306.11957 (2023)
10. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: Training deep neural networks with biased data. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9004–9012 (2019). https://doi.org/10.1109/CVPR.2019.00922
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015)
12. Kirichenko, P., Izmailov, P., Wilson, A.G.: Last layer re-training is sufficient for robustness to spurious correlations. ICLR (2023)
13. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision **128**(7), 1956–1981 (2020)

14. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
15. Li, Z., Evtimov, I., Gordo, A., Hazirbas, C., Hassner, T., Ferrer, C.C., Xu, C., Ibrahim, M.: A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20071–20082 (2023)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
17. Liu, e.a.: Discovering failure modes of text-guided diffusion models via adversarial search. In: ICLR (2023)
18. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just train twice: Improving group robustness without training group information. In: International Conference on Machine Learning. pp. 6781–6792. PMLR (2021)
19. Luccioni, A.S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Analyzing societal representations in diffusion models. arXiv preprint arXiv:2303.11408 (2023)
20. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
21. Marcus, G., Davis, E., Aaronson, S.: A very preliminary analysis of dall-e 2. arXiv preprint arXiv:2204.13807 (2022)
22. Meister, N., Zhao, D., Wang, A., Ramaswamy, V.V., Fong, R.C., Russakovsky, O.: Gender artifacts in visual datasets. ArXiv **abs/2206.09191** (2022)
23. Mondal, A.K., Singhal, L., Tiwary, P., Singla, P., Prathosh, A.P.: Minority oversampling for imbalanced data via class-preserving regularized auto-encoders. In: International Conference on Artificial Intelligence and Statistics (2023)
24. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
25. Qraitem, M., Saenko, K., Plummer, B.A.: Bias mimicking: A simple sampling approach for bias mitigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20311–20320 (2023)
26. Ramaswamy, V.V., Kim, S.S.Y., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: CVPR (2021)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
28. Ryu, H.J., Mitchell, M., Adam, H.: Improving smiling detection with race and gender diversity. CoRR **abs/1712.00193** (2017)
29. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: ICLR (2020)
30. 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. In: International Conference on Learning Representations (ICLR) (2020)
31. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-

to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
32. Sariyildiz, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
33. Sharmanska, V., Hendricks, L.A., Darrell, T., Quadrianto, N.: Contrastive examples for addressing the tyranny of the majority (2020)
34. Tartaglione, E., Barbano, C.A., Grangetto, M.: End: Entangling and disentangling deep representations for bias correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13508–13517 (June 2021)
35. Wang, A., Narayanan, A., Russakovsky, O.: REVISE: A tool for measuring and mitigating bias in image datasets. In: European Conference on Computer Vision (ECCV) (2020)
36. Wang, X., Lyu, Y., Jing, L.: Deep generative model for robust imbalance classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14124–14133 (2020)
37. Wang, Z., Qinami, K., Karakozis, I., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
38. Zhang, M., Sohoni, N.S., Zhang, H.R., Finn, C., Ré, C.: Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. arXiv preprint arXiv:2203.01517 (2022)
39. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: CVPR (2017)
40. Zhao, D., Wang, A., Russakovsky, O.: Understanding and evaluating racial biases in image captioning. In: International Conference on Computer Vision (ICCV) (2021)