DiffInject: Revisiting Debias via Synthetic Data Generation using Diffusion-based Style Injection

Donggeun Ko*
Aim Future
sean.ko@aimfuture.ai

Sangwoo Jo*
Minds and Company
sangwoo.jo@mnc.ai

Dongjun Lee
Maum AI
akm5825@maum.ai

Namjun Park, Jaekwang Kim†
Convergence Program of Social Innovation, Sungkyunkwan University
{951011jun, linux}@skku.edu

Abstract

Dataset bias is a significant challenge in machine learning, where specific attributes, such as texture or color of the images are unintentionally learned resulting in detrimental performance. To address this, previous efforts have focused on debiasing models either by developing novel debiasing algorithms or by generating synthetic data to mitigate the prevalent dataset biases. However, generative approaches to date have largely relied on using bias-specific samples from the dataset, which are typically too scarce. In this work, we propose, DiffInject, a straightforward yet powerful method to augment synthetic bias-conflict samples using a pretrained diffusion model. This approach significantly advances the use of diffusion models for debiasing purposes by manipulating the latent space. Our framework does not require any explicit knowledge of the bias types or labelling, making it a fully unsupervised setting for debiasing. Our methodology demonstrates substantial result in effectively reducing dataset bias.

1. Introduction

Deep learning networks and algorithms often inadvertently learn biases from extensive benchmark datasets. These biases, such as textures or colors, enable models to adopt shortcuts, leading to incorrect image classification. For instance, in a scenario where the majority of images depicting alligators are set against backgrounds of rivers or ponds, with scant examples of alligators on land, deep learning classifiers may rely on background features (e.g., river equates to alligator, land equates to horse) as ”shortcuts” for classification. In this context, we classify the background (river or land) as *task-irrelevant* features, whereas the subject of the image (the alligator) is a *task-relevant* feature. Features such as skin tones, genders, and colors also constitute *task-irrelevant* features that can obstruct the classifiers' ability to accurately represent objects within an image.

Prior approaches to reducing bias have explored supervised learning techniques, which depend on the annotation of biases or labels. An alternative strategy involves unsupervised learning, where generative models are employed to enhance the biased data without prior knowledge of the biases. This presents a significant advantage, as it suggests that bias mitigation should be performed without direct human supervision.

Diffusion-based models [6, 10, 24], have demonstrated superior performance compared to GANs [9, 13] in generating synthetic images. In particular, variants of Stable Diffusion [26] have achieved remarkable success in producing high-quality synthetic images. Recent research [3, 7, 30, 33] has shown that synthetic datasets generated by these models can significantly contribute to the learning or enhancement of visual representations in deep learning models. Consequently, leveraging generative models to translate bias-conflict features and augment data presents a promising avenue for enabling biased image classifiers to accurately learn and represent bias-conflict features.

In this paper, we propose a novel framework, DiffInject, where we “inject” or translate bias-conflict features into the data sample and generate synthetic dataset via leveraging the diffusion model. Our framework is composed of four major steps: 1) Overfit an image classifier into the biased dataset and extract samples with top-K loss in which we assume are bias-conflict images, 2) Train a diffusion model using a pronounced image benchmark dataset that generally encompasses the domain of the biased benchmark dataset, 3) Translate or inject the bias-conflict features content by leveraging the *h-space* of the diffusion model and translating the feature into the original bias-aligned image, 4) De-

[*] Equal Contribution †Corresponding author

bias the biased-classifier with the augmented dataset. Injecting the content of the top-K loss samples will allow the data to be translated and follow the distribution of bias-conflict images, allowing the biased model to capture the task-relevant features.

To the best of our knowledge, we believe our method is the first to explore leveraging the diffusion model in model debiasing via unsupervised learning. Our extensive experiments demonstrate that style injection of top-K loss samples to the original biased dataset allows to learn visual representation of biased-conflict samples extensively, thus debiasing the classifier.

2. Method

2.1. Preliminaries

Extracting Samples with High Losses. To extract bias-conflict samples from the dataset, we select top-K loss samples, where K denotes number of samples with top loss values, by intentionally overfitting the classifier where we named as bias classifier, f_B. By utilizing generalized cross-entropy (GCE) loss [32], we train the classifier from scratch to maximize the representation to become biased by prioritizing samples in the dataset that are “easy-to-learn” [2, 23] with high probability values which we believe are *bias-align* samples. Thus, GCE loss is formulated as follows:

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

where $p(x;\theta),y)$ denotes softmax output of the classifier, $p_y(x;\theta)$ is the probability of the target attribute y and $q \in (0,1]$ is a hyperparameter which controls the strength of amplification to make the model “easy-to-learn”. Classifier's parameters are denoted as θ. The GCE loss allows the classifier to gradually become biased by maximizing the weights on the gradients of the samples with higher probability p_y, formulated as follows:

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

Overfitting the Classifier to become Biased. This allows the model have higher GCE losses and maximizes the loss when the model sees *bias-conflict* samples, eventually learning shortcuts provided by abundant bias-align samples in the dataset.

After overfitting the classifier to become biased, we are able to select samples with top-K losses by calculating cross-entropy loss of all the train data output. We chose K as 10 following the method of AmpliBias [16] and for fair comparison with other generative model methods as well. Selection process of the top-K loss samples from the overfitted classifier is as follows:

$$x_{\mathrm{c}} = \operatorname*{argmax}_{x_i \in D_{\text{biased}}} CE(f_B(x_i), y_i), \quad \text{where } x_{\mathrm{c}} \in \mathcal{X}_{\mathrm{c}}, \tag{3}$$

$$\begin{aligned}\mathcal{X}_c = \{x_{c1}, x_{c2}, \ldots, x_{cK}\}, &\text{where}\\ L(f_B(x_{c1})) > L(f_B(x_{c2})) > \ldots &> L(f_B(x_{cK}))\end{aligned} \tag{4}$$

where x_c, x_{cK} and $\mathcal{X}_c$ denotes the extracted bias-conflict samples, K-th bias-conflict sample and a set of bias-conflict samples, respectively. $L(f_B(x_{c1})) > L(f_B(x_{c2}))$ signifies that the loss of x_{c1} is greater than x_{c2}, and top-K losses are ordered from highest to lowest in $\mathcal{X}_c$.

2.2. Diffusion Model with P2 Weighting

We train a diffusion model with Perception Prioritized (P2) weighting objective $L_{P2} = \sum_t \lambda_t^{P2} L_t$ from a standard benchmark dataset, where the loss function for each time step t is defined as the following [5].

$$\begin{aligned}L_t &= D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t))\\ &= \mathbb{E}_{x_0,\epsilon}[\lambda_t^{P2}||\epsilon - \epsilon_\theta(x_t,t)||^2]\end{aligned} \tag{5}$$

with the weighting scheme λ_t^{P2} defined as

$$\lambda_t^{P2} = \frac{\lambda_t}{(k + \text{SNR}(t))^\gamma}, \tag{6}$$

where $\text{SNR}(t) = \alpha_t/(1-\alpha_t), \alpha_t = \prod_{s=1}^t (1-\beta_s)$ from standard diffusion model notation [6, 10, 24]. Both hyperparameters γ and k are set as 1, where γ controls the degree of learning perceptually rich contents and k determines the sharpness of the weighting scheme. We compare the quality of generated images with baseline diffusion models [10, 24] and further validate the use of P2 weighting as our model for synthetic data generation.

2.3. Injecting Biased Contents

In our approach, we utilize images with high loss values to generate synthetic bias-conflict samples by integrating their biased content into randomly chosen bias-aligned samples. To achieve this, we manipulate the bottleneck layer of the U-Net architecture, denoted as *h-space* [19], as a method of content injection during the DDIM reverse process proposed in InjectFusion [12]:

$$x_{t-1} = \sqrt{\alpha_{t-1}} P_t(\epsilon_t^\theta(x_t|\tilde{h}_t)) + D_t(\epsilon_t^\theta(x_t)) + \sigma_t z_t \tag{7}$$

where $P_t(\epsilon_t^\theta(x_t)|\tilde{h}_t)$ denotes the predicted x_0, $D_t(\epsilon_t^\theta(x_t))$ denotes the direction pointing to x_t, and $\tilde{h}_t$ is the modified *h-space* replacing the original h_t. The bottleneck layer is modified using normalized spherical interpolation (Slerp)

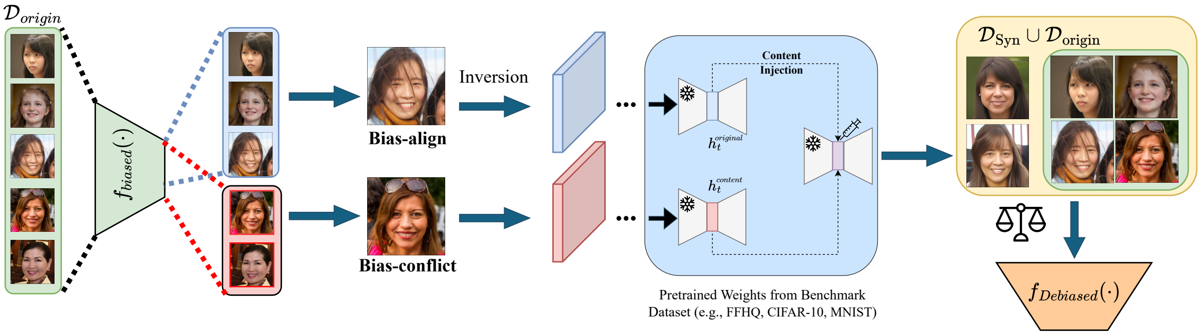

Figure 1. Overall framework of our proposed method, DiffInject.

between h_t's which proves to exhibit fewer artifacts when compared to either replacing or adding them:

$$\tilde{h}_t = f(h_t^{\text{original}}, h_t^{\text{content}}, \gamma) = \text{Slerp}\left(h_t^{\text{original}}, \frac{h_t^{\text{content}}}{\|h_t^{\text{content}}\|} \cdot \|h_t\|, \gamma\right) \quad (8)$$

where h_t^{original} and h_t^{content} is the *h-space* of the original and content image respectively, and $\gamma \in [0, 1]$ denotes content injection ratio.

We apply content injection at both global and local levels. Local content injection is performed by masking the targeted area of the *h-space* before applying Slerp, with the resulting interpolated h_t subsequently inserted back into the original feature map. This image editing process occurs during the early stage of the generative process $[T, t_{\text{edit}}]$ followed by the stochastic noise injection during interval $[t_{\text{boost}}, 0]$. Then, we generate synthetic data samples to represent a certain proportion of the overall dataset, with the term "bias-conflict ratio" used throughout this paper. Further details are provided in Appendix B.

2.4. Training Unbiased Classifier

Following the synthetic data generation method described in Section 2.3, we construct an unbiased dataset D_{syn}. Subsequently, we mitigate the bias in our biased classifier by training on a combined dataset comprising both synthetic and original data, denoted as $D_{total} = D_{syn} \cup D_{orig}$. This enables the model to learn more general visual representations of task-relevant features within in the dataset, thereby enhancing the debiasing of the learning process. It is important to note that labels for synthetic data is automatically assigned based on the bias-conflicted samples from which they were generated.

3. Experiments

Datasets We conduct experiments on four datasets with their matching class and bias attributes. Details are as follows: **Colored MNIST**: (Number-Color), **Corrupted CIFAR-10**: (Object-Noise), **BFFHQ**: (Age-Gender), **Dogs & Cats**: (Animal-Fur Color).

Baselines We compare our method with vanilla network, LfF, DisEnt, BiasEnsemble, A^2, and AmpliBias as our baselines. Vanilla network is defined as multi-layer perceptron (MLP) with three hidden layers for Colored MNIST, and ResNet-18 for the remaining datasets.

Implementation Details We pretrained ADM [6] with P2-weighting on four widely recognized computer vision benchmark datasets: MNIST, CIFAR-10, FFHQ, and AFHQ. Subsequently, we use the pretrained model weights to apply InjectFusion on our benchmark datasets: Colored MNIST (CMNIST), Corrupted CIFAR-10 (CCIFAR-10), BFFHQ, and Dogs & Cats. Further implementation details are described in Appendix B.

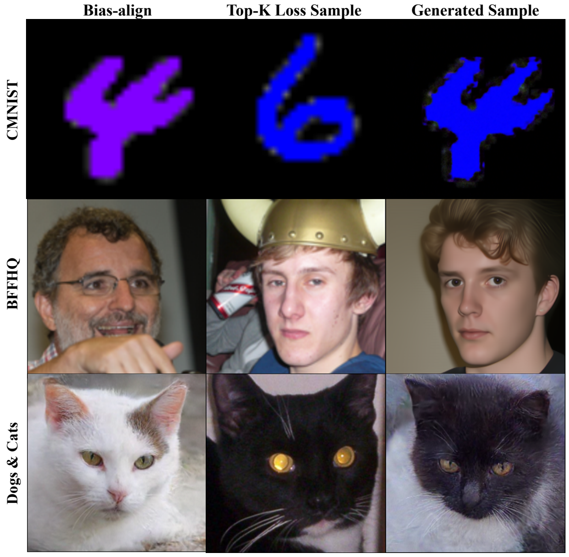

Figure 2. Generated bias-conflict samples with DiffInject. The three columns represent samples from the original dataset, top-k loss samples and generated samples, respectively.

Table 1. Performance in accuracy (%) for unbiased test sets across four benchmark datasets with varying ratios of bias-conflicting samples. Best performance is highlighted in bold and second-best is underlined, respectively.

Methods	Synthetic				Real-World			
	Colored MNIST		Corrupted CIFAR-10		BFFHQ		Dogs & Cats	
	1.0%	5.0%	1.0%	5.0%	1.0%	5.0%	1.0%	5.0%
Vanilla	32.59	82.44	23.62	41.68	60.68	83.36	73.01	83.51
LfF	74.23	85.33	33.57	49.47	69.89	78.31	69.92	82.92
DisEnt	78.89	89.60	36.49	51.88	66.00	80.68	69.49	82.78
LfF + BE	81.17	90.04	34.77	52.16	75.08	85.48	81.52	88.60
A^2	83.92	91.64	27.54	37.60	78.98	86.22	71.15	83.07
AmpliBias	67.79	74.88	**45.95**	**52.22**	81.80	87.34	73.30	84.67
DiffInject	**85.58**	**92.29**	28.24	33.11	**81.90**	**89.90**	**82.35**	**93.60**

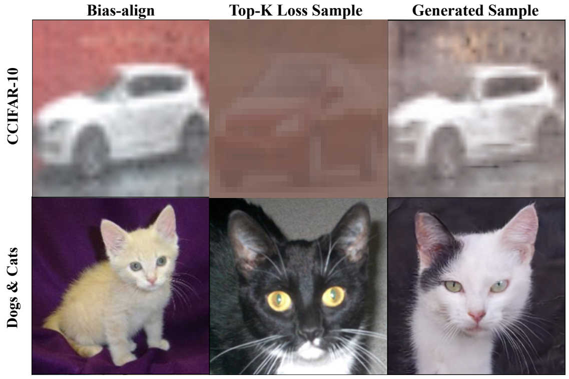

Figure 3. Generated bias-conflict samples with artifacts from our framework, DiffInject.

3.1. Analysis

Quantitative Results Table 1 presents image classification accuracies on the unbiased test sets of synthetic datasets and real world datasets, where the ratio of bias-conflicting samples is varied at 1% and 5%. Most notably, DiffInject outperforms baseline methods such as LfF [23] and DisEnt [20] on real world datasets by a substantial margin. DiffInject also achieves state-of-the-art performance on CMNIST, but lacks performance in CCIFAR-10. It is important to note that our method does not require prior knowledge in bias types and manual labeling of synthetic data, yet demonstrating superior performance across most benchmark datasets compared to the baselines.

Qualitative Results We analyze the quality of synthetic samples generated from DiffInject, as illustrated in Fig. 2. The three columns depict bias-aligned samples from the original dataset, original samples with top-K loss, and well-generated bias-conflict samples via DiffInject, respectively. Synthetic data generated with DiffInject provide higher quality and realism, facilitating the model to learn debiased representation. For instance, the generated sample in BFFHQ effectively captures the bias-conflict attribute of "young-male", enriching the visual features of the dataset. Successful injection of bias-conflict attributes is also demonstrated in CMNIST and Dogs & Cats datasets.

Limitations Figure 3 demonstrates synthetic samples generated with artifacts. To ensure fair comparison and prevent selective sampling of synthetic data, we included bias-conflict data with artifacts in D_{syn} for training the classifier. For instance, in the case of Dogs & Cats, the generated sample with artifacts fails to transfer the dark fur color observed in the bias-conflict samples. However, the presence of artifact may have positively contributed to the debiasing process, as the mixed fur color (black & white) enables the model to learn diverse visual representation. For CCIFAR-10, our generated sample fails to effectively capture the bias attributes, which may have negatively impacted the model performance as demonstrated in Table 1. Further works can be explored to successfully extract and inject texture bias attributes presented in CCIFAR-10.

4. Conclusion

We propose, DiffInject, a novel framework for debiasing the image classifier by augmenting synthetic data through semantic manipulation of the latent space within the diffusion model. Our approach eliminates the need for manual labeling of synthetic data and explicit knowledge of bias types in the samples, yet generating high quality bias-conflict synthetic samples. With our results demonstrating significant performance increase over benchmark datasets, we believe our work inspires the future work of leveraging diffusion models in model debiasing.

Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-RS-2023-00259497) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00254129, No.RS-2023-00230561, Development of Conversational Integrated AI Search Engine based on Multimodal Technology), Graduate School of Metaverse Convergence (Sungkyunkwan University)), Aim Future, Minds and Company, and PseudoLab.

References

[1] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. *arXiv preprint arXiv:2309.06933*, 2023. 1

[2] Jaeju An, Taejune Kim, Donggeun Ko, Sangyup Lee, and Simon S Woo. Aˆ 2: Adaptive augmentation for effectively mitigating dataset bias. In *Proceedings of the Asian Conference on Computer Vision*, pages 4077–4092, 2022. 2, 1

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 1

[4] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 1

[5] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 2, 1

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2, 3

[7] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. *arXiv preprint arXiv:2312.04567*, 2023. 1

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1

[12] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free content injection using h-space in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5151–5161, 2024. 2, 1

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[14] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019. 1

[15] Kangyeol Kim, Sunghyun Park, Junsoo Lee, and Jaegul Choo. Reference-based image composition with sketch via structure-aware diffusion model. *arXiv preprint arXiv:2304.09748*, 2023. 1

[16] Donggeun Ko, Dongjun Lee, Namjun Park, Kyoungrae Noh, Hyeonjin Park, and Jaekwang Kim. Amplibias: Mitigating dataset bias through bias amplification in few-shot learning for generative models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4028–4032, 2023. 2, 1

[17] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1

[18] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 1

[19] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2

[20] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021. 4, 1

[21] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 1

[22] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 1

[23] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 2, 4, 1

[24] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1, 2

[25] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021. 1

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[28] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 1

[29] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13508–13517, 2021. 1

[30] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In *NeurIPS*, 2023. 1

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[32] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 2

[33] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Improving classification accuracy on arbitrary datasets using synthetic data. 2023. 1