Inclusive GAN: Improving Data and Minority Coverage in Generative Models

Ning Yu[1,2], Ke Li[3,5,6], Peng Zhou[1]
Jitendra Malik[3], Larry Davis[1], and Mario Fritz[4]

[1] University of Maryland, College Park, United States
[2] Max Planck Institute for Informatics, Saarbrücken, Germany
[3] University of California, Berkeley, United States
[4] CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
[5] Institute for Advanced Study, Princeton, United States
[6] Google, Seattle, United States
ningyu@mpi-inf.mpg.de ke.li@eecs.berkeley.edu pengzhou@cs.umd.edu
malik@eecs.berkeley.edu lsd@cs.umd.edu fritz@cispa.saarland

Abstract. Generative Adversarial Networks (GANs) have brought about rapid progress towards generating photorealistic images. Yet the equitable allocation of their modeling capacity among subgroups has received less attention, which could lead to potential biases against underrepresented minorities if left uncontrolled. In this work, we first formalize the problem of minority inclusion as one of data coverage, and then propose to improve data coverage by harmonizing adversarial training with reconstructive generation. The experiments show that our method outperforms the existing state-of-the-art methods in terms of data coverage on both seen and unseen data. We develop an extension that allows explicit control over the minority subgroups that the model should ensure to include, and validate its effectiveness at little compromise from the overall performance on the entire dataset. Code, models, and supplemental videos are available at GitHub.

Keywords: GAN, Minority Inclusion, Data Coverage

1 Introduction

Photorealistic image generation has increasingly become reality, thanks to the emergence of large-scale datasets [10,31,47] and deep generative models [25,15,26,29]. However, these advances have come at a cost: there could be potential biases in the learned model against underrepresented data subgroups [45,53,41,16,17]. The biases are rooted in the inevitable imbalance in the dataset [38], which are preserved or even exacerbated by the generative models [53]. In particular, reconstructive (non-adversarial) generative models like variational autoencoders (VAEs) [25,36] can preserve data biases against minorities due to their objective of reproducing the frequencies images occur in the dataset, while adversarial generative models (GANs) [15,12,11] can implicitly disregard infrequent images

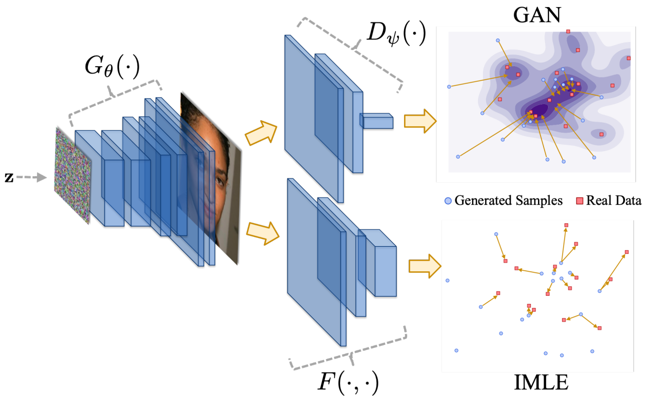

Fig. 1. The diagram of our method. It harmonizes adversarial (GAN) and reconstructive (IMLE) training in one framework without introducing an auxiliary encoder. GAN guides arbitrary sampling towards generating realistic appearances approximate to some real data while IMLE ensures data coverage where there are always generated samples approximate to each real data. See Section 3.3 for more details where G_θ and D_ψ represent the trainable generator and discriminator in a GAN, and F represents a distant metric, in some cases, a pre-trained neural network.

due to the well-established problem of mode collapse [42,29], thereby further introducing model biases on top of data biases. This issue is particularly acute from the perspective of minority inclusion, because training data associated with minority subgroups by definition do not form dominant modes. Consequently, data from minority groups are rare to begin with, and would not be capable of being produced by the generative model at all due to mode collapse.

In this work, we aim to improve the *comprehensive* performance of the state-of-the-art generative models, with a specific focus on their coverage of minority subgroups. We start with an empirical study on the correlation between data biases and model biases, and then formalize the objective of alleviating model bias in terms of improving data coverage, in particular over the minority subgroups. We propose a new method known as IMLE-GAN that achieves competitive image quality while ensuring improved coverage of minority groups.

Our method harmonizes adversarial and reconstructive generative models, in the process combining the benefits of both. Adversarial models have evolved to generate photorealistic results, whereas reconstructive models offer guarantees on data coverage. We build upon one of the state-of-the-art implementations of adversarial models, i.e., StyleGAN2 [23], and incorporate it with the Implicit Maximum Likelihood Estimation (IMLE) framework [29], which is at its core reconstructive. See Figure 1 for a diagram.

Different from the existing hybrid generative models [26,42,37,4] that require training an auxiliary encoder network alongside a vanilla GAN, our method operates purely with the standard components of a GAN. This brings two main benefits: (1) it sidesteps the complication from combining the minimax objective used by adversarial models and the pure minimization objective used by reconstructive models, and (2) it avoids carrying over the practical issues of training auxiliary encoder, like posterior collapse [5,24], which can cause the regression-to-the-mean problem, leading to blurry images.

We validate our method with thorough experiments and demonstrate more comprehensive data coverage that goes beyond that of existing state-of-the-art methods. In addition, our method can be flexibly adapted to ensure the inclusion of specified minority subgroups, which cannot be easily achieved in the context of existing methods.

Contributions. We summarize our main contributions as follows: (1) we study the problem of underrepresented minority inclusion and formalize it as a data coverage problem in generative modeling; (2) we present a novel paradigm of harmonizing adversarial and reconstructive modeling for improving data coverage; (3) our experiments set up a new suite of state-of-the-art performance in terms of covering both seen and unseen data; and (4) we develop an effective extension of our technique to ensure inclusion of the specified minority subgroups.

2 Related Work

Bias mitigation efforts for machine learning. Bias in machine learning results from data imbalance, which can be detected and alleviated by three categories of approaches: The pre-process approaches that purify data from bias before training [7,13,14,49], the in-process approaches that enforce fairness during training with constraints or regularization in the objectives [22,48,38,51], and the post-process approaches that adjust the output from a learned model [21,19]. A comprehensive survey [32] articulates this taxonomy. These approaches target biases in classification and cannot be adapted to generative modeling.

Bias mitigation efforts for generative models. There have been relatively few papers [45,53,41,16,17] that focus on biases in generative models. [45,41,16], motivated from benefiting a downstream classifier, mainly aim for fair generation conditioned on attribute inputs, in terms of yielding allocative decisions and/or removing the correlation between generation and attribute conditions. [53] focuses on understanding the inductive bias so as to investigate the generalization of generative models. [17] proposes an importance weighting strategy to compensate for the biases of learned generative models. Different from their goals and solutions that equalize performance across different data subgroups possibly at the cost of overall performance, we instead aim to improve the overall data coverage, with a specific purpose of ensuring more significant gains over the underrepresented minorities.

Data coverage in GANs. GANs are finicky to train because of the minimax formulation and the alternating gradient ascent-descent. In addition, GANs are known to exhibit mode collapse, where the generator only learns to generate a subset of the modes of the underlying data distribution. To alleviate mode collapse in GANs, some methods propose to improve the minimax loss function [33,2,18,1], some methods apply constraints or regularization terms along with the minimax objectives [9,3,43,46,30], and some other methods aim to modify the discriminator designs [44,52,34,35]. These directions are orthogonal to our research while, in principle, demonstrate less effective data coverage than the hybrid models below.

Data coverage in hybrid generative models. Reconstructive (non-adversarial) generative models like variational autoencoders (VAEs) [25,36], on the other hand, are more successful at data coverage because they explicitly try to maximize a lower bound on the likelihood of the real data. This motivates a variety of designs for hybrid models that combine reconstruction and adversarial training. α-GAN [37] is trained to reconstruct pixels while VAEGAN [26] is trained to reconstruct discriminator features. ALI [12], BiGAN [11], and SVAE [8] propose to instead jointly match the bidirectional mappings between data and latent distributions. VEEGAN [42] is designed with reconstruction in the latent space, in the purpose of avoiding the metric dilemma in the data space. Hybrid models benefit for mode coverage, but deteriorate generation fidelity in practice, because of their dependency on auxiliary encoder networks. In contrast, our method follows the idea of hybrid models, but avoids an encoder network and instead apply all training back-propagation through the generator. A recent non-adversarial generative framework, Implicit Maximum Likelihood Estimation (IMLE) [29], satisfies our design. We discuss more about the advantages of IMLE in Section 3.2.

3 Inclusive GAN for Data and Minority Coverage

Our method is a novel paradigm of harmonizing the strengths of adversarial (Section 3.1) and reconstructive generative models (Section 3.2) that avoids mode collapse. The harmonization efforts (Section 3.3) are necessary and non-trivial due to the incompatibility between the two, which is validated in the supplementary material. In Section 3.4 we show the straightforward adaptation of our method to improve minority inclusion.

3.1 Adversarial Generation: GANs

Photorealistic image generation can be viewed as the problem of sampling from the unknown probability distribution of real-world images. Generative Adversarial Networks (GANs) [15] introduce an elegant solution for distribution estimation, which is formulated as a discriminative classification problem, and enables supervised learning methods to be used for this task.

A GAN consists of two deep neural networks: a generator $G_\theta : \mathbb{R}^d \mapsto \mathbb{R}^D$ and a discriminator $D_\psi : \mathbb{R}^D \mapsto [0,1]$. The generator maps a latent noise vector $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ to an image, and the discriminator predicts the probability that the image it sees is real. The real ground truth images are denoted as $\mathbf{x} \sim \hat{p}(\mathbf{x})$, sampled from an unknown distribution $\hat{p}(\mathbf{x})$. The discriminator is trained to maximize classification accuracy while the generator is trained to produce images that can fool the discriminator. More precisely, the objective is shown in Eq. 1:

$$\min_{\theta} \max_{\psi} L^{adv}(\theta, \psi) = \mathbb{E}_{\mathbf{x} \sim \hat{p}(\mathbf{x})} \left[\log D_{\psi}(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)} \left[\log(1 - D_{\psi}(G_{\theta}(\mathbf{z})))\right] \tag{1}$$

Unfortunately, GANs are unstable to train and suffer from mode collapse: While each generated sample gets to pick a mode it is drawn to, each mode does not get to pick a generated sample. After training, the generator will not be able to generate samples around the "unpopular" modes.

Minority modes are precisely the "unpopular" modes that are more likely to be collapsed. As shown in Section 4.3 and Figure 2, minority subgroups with diverse appearances indeed bring more challenges to generative modeling and are allocated worse coverage compared to the others. Therefore, we propose to leverage reconstructive models to improve the coverage of minority subgroups.

3.2 Reconstructive Generation: IMLE

Our novel paradigm is based on a recent reconstructive framework, Implicit Maximum Likelihood Estimation (IMLE) [29], that favors complete mode coverage. IMLE avoids mode collapse by reversing the direction in which generated samples are matched to real modes. In GANs, each generated sample is effectively matched to a real mode. In IMLE, each real mode is matched to a generated sample. This ensures that all real modes, including each underrepresented minority mode, are matched, and no real mode is left out.

Mathematically, IMLE tackles the optimization problem in Eq. 2:

$$\begin{align} & \min_{\theta} \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)} \left[\mathbb{E}_{\mathbf{x} \sim \hat{p}(\mathbf{x})} \left[\min_{i \in \{1, \ldots, m\}} \|G_{\theta}(\mathbf{z}_i) - \mathbf{x}\|_2^2 \right] \right] \tag{2} \\ = & \min_{\theta} \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)} \left[\mathbb{E}_{\mathbf{x} \sim \hat{p}(\mathbf{x})} \left[\|G_{\theta}(\mathbf{z}^*(\mathbf{x})) - \mathbf{x}\|_2^2 \right] \right], \tag{3} \\ \text{where } \mathbf{z}^* = & \operatorname*{argmin}_{i \in \{1, \ldots, m\}} \|G_{\theta}(\mathbf{z}_i) - \mathbf{x}\|_2^2 \tag{4} \end{align}$$

The joint optimization is achieved by alternating between the two decoupled phases until convergence. The first phase corresponds to the inner optimization, where we search for each $\mathbf{x}$ the optimal $\mathbf{z}^*(\mathbf{x})$ from the latent vector candidates, given a fixed G_{θ}. This is implemented by the Prioritized DCI [28], a fast nearest neighbor search algorithm. The second phase corresponds to the outer optimization, where we train the generator in the regular back-propagation manner, given pairs of $(\mathbf{x}, \mathbf{z}^*(\mathbf{x}))$.

One significant advantage of IMLE over the other reconstructive models is the elimination of the need for an auxiliary encoder. The encoder encourages mode coverage but at the cost of either deviating the latent sampling distribution from the original prior (in VAEGAN [26]) or absorbing the training gradients before substantially back-propagating to the generator (in VEEGAN [42]). Unlike them, IMLE directly samples latent vector from a natural prior during training and encourages explicit reconstruction fully upon the generator.

3.3 Harmonizing Adversarial and Reconstructive Generation: IMLE-GAN

Below we propose a way to harmonize adversarial training with the IMLE framework, so as to ensure both generation quality (precision) and coverage (recall) simultaneously.

The vanilla hybrid model between IMLE and GAN is to directly add the adversarial loss in Eq. 1 to the non-adversarial loss in Eq. 2. This has two problems because of (1) differences in the domains over which latent vectors are sampled and (2) differences in the metric spaces on which GAN and IMLE operate. For (1), in the case of GAN, a different latent vector is randomly sampled every iteration, whereas in the case of IMLE, many latent vectors are sampled at once (over which matching is performed) and are kept fixed for many iterations. The former gives up control over which data point each latent vector is asked to generate by the discriminator, but can avoid overfitting to any one latent vector. The latter explicitly controls which latent vectors are matched to data points, but can overfit to the set of matched latent vectors until they are resampled. For (2), in the case of GAN, the discriminator takes the inner product between the features and the weight vector of the last layer to produce a realism score, and so it effectively operates on features of images; on the other hand, in the case of IMLE, matching is performed on raw pixels.

To bridge the gap in losses, we propose two adaptations that better harmonize the GAN and IMLE objectives. First, to make the domain over which latent vectors are sampled denser, we augment the matched latent vectors with random linear interpolations. Second, to make the spaces on which the two losses are computed more comparable, we measure the reconstruction loss in a deep feature space instead of pixel space, such that it contains a comparable amount and level of semantic information to that used by the discriminator. Mathematically, our goal is to optimize Eq. 5:

$$\min_{\theta}\max_{\psi} L^{adv}(\theta,\psi) + \mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_m\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\lambda L^{rec}(\theta) + \beta L^{itp}(\theta)\right] \tag{5}$$

Here $L^{adv}(\theta,\psi)$ is as defined in Eq. 1,

$$\begin{aligned}
L^{rec}(\theta) =& \mathbb{E}_{\mathbf{x}\sim\hat{p}(\mathbf{x})}\left[\|F(G_\theta(\mathbf{z}^*(\mathbf{x}))) - F(\mathbf{x})\|_2^2\right] && (6)\\
\text{where } \mathbf{z}^*(\mathbf{x}) =& \operatorname*{argmin}_{i\in\{1,\ldots,m\}} \|F(G_\theta(\mathbf{z}_i)) - F(\mathbf{x})\|_2^2, && (7)\\
\text{and } L^{itp}(\theta) =& \mathbb{E}_{\mathbf{x},\widetilde{\mathbf{x}}\sim\hat{p}(\mathbf{x}),\alpha\sim U[0,1]}\left[\alpha\|F(G_\theta(\mathbf{z}^*(\alpha,\mathbf{x},\widetilde{\mathbf{x}}))) - F(\mathbf{x})\|_2^2 + \right. && (8)\\
& \left.(1-\alpha)\|F(G_\theta(\mathbf{z}^*(\alpha,\mathbf{x},\widetilde{\mathbf{x}}))) - F(\widetilde{\mathbf{x}})\|_2^2\right] && (9)\\
\text{where } \mathbf{z}^*(\alpha,\mathbf{x},\widetilde{\mathbf{x}}) =& \alpha\mathbf{z}^*(\mathbf{x}) + (1-\alpha)\mathbf{z}^*(\widetilde{\mathbf{x}}) && (10)
\end{aligned}$$

Here Eq. 6 generalizes Eq. 3 by computing distance in feature space, where $F(\cdot)$ is a fixed function to compute features of images. Eq. 8 and 9 defines the interpolation loss, which linearly interpolates between two matched latent vectors $\mathbf{z}^*(\mathbf{x}), \mathbf{z}^*(\widetilde{\mathbf{x}})$ (as shown in Eq. 10) and tries to make the image generated from the interpolated latent vector $\mathbf{z}^*(\alpha,\mathbf{x},\widetilde{\mathbf{x}})$ similar to the two ground truth

Algorithm 1: IMLE-GAN with Minority Inclusion

```
Data: Real training data $\hat{p}(\mathbf{x})$ and a specified minority subgroup $\hat{q}(\mathbf{y})$
Result: A generator $G_\theta$ with specified minority inclusion performance
for epoch $= \{1, \ldots, E\}$ do
    if epoch $\% \ S == 0$ then
        Sample $\mathbf{z}_1, \ldots, \mathbf{z}_m \sim \mathcal{N}(0, \mathbf{I}_d)$ i.i.d.;
        for $\mathbf{y}_j \sim \hat{q}(\mathbf{y})$ do
            $\mathbf{z}^*(\mathbf{y}_j) \leftarrow \arg\min_{i \in \{1,\ldots,m\}} F(G_\theta(\mathbf{z}_i), \mathbf{y}_j)$;
        end
    end
    for $\mathbf{x}_k \sim \hat{p}(\mathbf{x})$ and $\mathbf{y}_i, \mathbf{y}_j \sim \hat{q}(\mathbf{y})$ do
        Sample $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_d)$;
        $L^{adv} \leftarrow \log D_\psi(\mathbf{x}_k) + \log(1 - D_\psi(G_\theta(\mathbf{z})))$;
        Sample $\delta_i, \delta_j \sim \mathcal{N}(0, \sigma \mathbf{I}_d)$ i.i.d.;
        $\mathbf{z}_i^* \leftarrow \mathbf{z}^*(\mathbf{y}_i) + \delta_i$;
        $\mathbf{z}_j^* \leftarrow \mathbf{z}^*(\mathbf{y}_j) + \delta_j$;
        $L^{rec} \leftarrow \frac{1}{2}(F(G_\theta(\mathbf{z}_i^*), \mathbf{y}_i) + F(G_\theta(\mathbf{z}_j^*), \mathbf{y}_j))$;
        Sample $\alpha \sim U[0, 1]$;
        $\mathbf{z}_{ij}^* = \alpha \mathbf{z}_i^* + (1 - \alpha)\mathbf{z}_j^*$;
        $L^{itp} \leftarrow \alpha F(G_\theta(\mathbf{z}_{ij}^*), \mathbf{y}_i) + (1 - \alpha) F(G_\theta(\mathbf{z}_{ij}^*), \mathbf{y}_j)$;
        $L \leftarrow L^{adv} + \lambda L^{rec} + \beta L^{itp}$;
        $\psi = \psi + \eta \nabla_\psi L$;
        $\theta = \theta - \eta \nabla_\theta L$;
    end
end
```

images $\mathbf{x}, \widetilde{\mathbf{x}}$ that correspond to the latent vectors at the endpoints. The weight on the distance to each ground truth image depends on how close the interpolated latent vector is to the endpoint, which is denoted by α. λ and β are used to balance each loss term. We experiment with four possible feature spaces: raw pixels, discriminator features [26], Inception features [40], and LPIPS features (i.e.: features such that the ℓ_2 distance between them is equivalent to the LPIPS perceptual metric [50]), and compare them in the the supplementary material.

3.4 Minority Coverage in IMLE-GAN

IMLE-GAN framework is designed to improve the overall mode coverage. One benefit compared to other hybrid models is that it is straightforward to adapt it for minority inclusion. We simply need to replace the empirical distribution over the entire dataset $\hat{p}(\mathbf{x})$ with a distribution $\hat{q}(\mathbf{x})$ whose support only covers a specified minority subgroup (i.e.: $\text{supp}(\hat{q}) \subset \text{supp}(\hat{p})$) in Eq. 6 and 8 (for reconstructive training) and leave Eq. 1 unchanged (for adversarial training). This ensures an explicit coverage over the minority while still carrying out the approximation to the entire real data. This comes with another advantage: because $\hat{q}(\mathbf{x})$ in practice has support over a much smaller set than $\hat{p}(\mathbf{x})$, there is less data imbalance and variance within the support of $\hat{q}(\mathbf{x})$ than in $\hat{p}(\mathbf{x})$, thereby requiring less model capacity to model. As a result, covering $\hat{q}(\mathbf{x})$ should be easier than covering $\hat{p}(\mathbf{x})$, and so the perceptual quality of samples tend to improve.

We summarize our IMLE-GAN algorithm with minority inclusion in Algorithm 1, where E is the number of training epochs, S indicates how often (in epochs) to update latent matching, m is the pool size of the latent vector candidates, δ_i, δ_j are the additive Gaussian perturbations, and η is the learning rate. We provide the hyperparameter settings in the supplementary material.

4 Experiments

We articulate the experimental setup in Section 4.1. In Section 4.2 we start with preliminary validation on Stacked MNIST dataset [33], an easy and interpretable task. In Section 4.3 we conduct empirical study to analyze the correlation between data bias and model bias. We then move on to the validation of our two harmonization strategies in the supplementary material. In Section 4.4 we perform comprehensive evaluation and comparisons on CelebA dataset [31], and finally specify minority inclusion applications in Section 4.5.

4.1 Setup

Datasets. For preliminary study, we employ Stacked MNIST dataset [33] for explicit data coverage evaluation. 240,000 RGB images in the size of 32×32 are synthesized by stacking three random digit images from MNIST [27] along the color channel, resulting in 1,000 explicit modes in a uniform distribution.

We conduct our main experiments on CelebA human face dataset [31], where the 40 binary facial attributes are used to specify minority subgroups. We sample the first 30,000 images in the size of 128×128 for GAN training, and sample the last 3,000 or 30,000 images for validation.

GAN backbone. We build our IMLE-GAN framework on the state-of-the-art StyleGAN2 [23] architecture for unconditional image generation. We reuse all their default settings.

Baseline methods. Besides the backbone StyleGAN2 [23], we also compare our method to eight techniques that show improvement in data coverage and/or generation diversity: SNGAN [34], Dist-GAN [43], DSGAN [46], PacGAN [30], ALI [12], VAEGAN [26], α-GAN [37], and VEEGAN [42]. For VAEGAN which originally involves image reconstruction in the discriminator feature space, we also experiment with three other distance metrics as discussed in Section 3.3. For fair comparisons, we replace the original architectures used in all methods with StyleGAN2. See supplementary material for their parameter settings.

Evaluation. For Stacked MNIST, following [33,42], we report the number of generated modes that is detected by a pre-trained mode classifier, as well as the KL divergence between the generated mode distribution and the uniform distribution. The statistics are calculated from 240,000 randomly generated samples.

Table 1. Comparisons on Stacked MNIST dataset. The statistics are calculated from 240,000 randomly generated samples. We indicate for each metric whether a higher (⇑) or lower (⇓) value is more desirable. We highlight the best performance in **bold**.

	# modes (max 1000) (⇑)	KL to uniform (⇓)
StyleGAN2 [23]	940	0.424
SNGAN [34]	571	1.382
DSGAN [46]	955	0.343
PacGAN [30]	908	0.638
ALI [12]	956	0.680
VAEGAN [26]	929	0.534
VEEGAN [42]	987	0.310
Ours LPIPS interp	**997**	**0.200**

For CelebA, Fréchet Inception Distance (FID) [20] is used to reflect both data quality (precision) and coverage (recall) in an entangled manner. We also explicitly measure the Precision and Recall [39] of a generative model w.r.t. the real dataset in the Inception space. Moreover, to emphasize on instance-level data coverage, we further include Inference via Optimization Measure (IvOM) [33] into our metric suite, which measures the mean retrieval error from a generative model given each query image. We also report the standard deviation of IvOM across 40 CelebA attributes, in order to evaluate the balance of generative coverage. For the generalization purpose, we evaluate over both the training set and a validation set (unseen during training). More details of the evaluation implementation are in the supplementary material.

4.2 Preliminary Study on Stacked MNIST

In a real-world data distribution, the notion of modes is difficult to quantize. We instead start with Stacked MNIST [33] where 1,000 discrete modes are unambiguously synthesized. This allows us to zoom in the challenge of mode collapse and facilitate a precise pre-validation.

We report the evaluation in Table 1. Our method narrows down the gap between experimental performance and the theoretical limit: It covers the most number of modes and achieves the closest mode distribution to the uniform distribution ground truth. This study validates the improved effectiveness of harmonizing IMLE with GAN, compared to the other GAN models or hybrid models, in terms of explicit mode/data coverage. This sheds the light and prequalifies to apply our method on more complicated real-world datasets.

4.3 Empirical Study on Data and Model Biases

As discussed in Section 2, data biases lead to biases in generative models. Even worse, a model without attention to minorities can exacerbate such biases against allocating adequate representation capacities to them. In this empirical study,

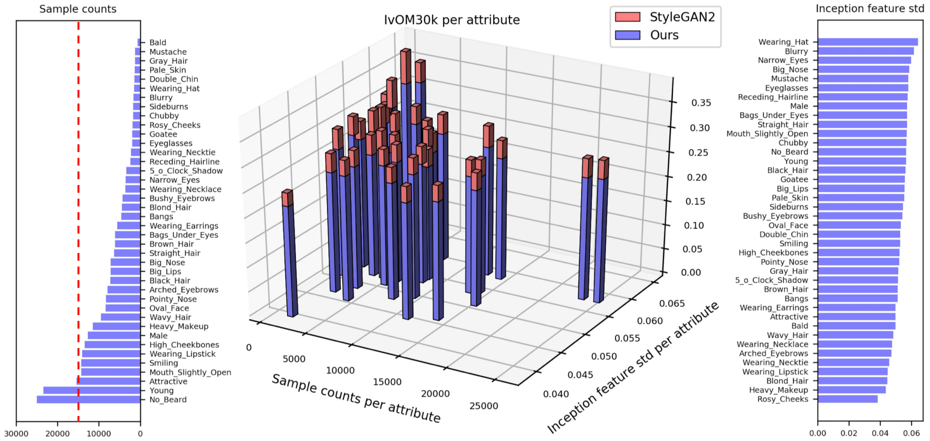

Fig. 2. Visualizations for data and model biases. Left: Sorted CelebA attribute histogram with a balance point marked by the red dashed line. Right: Sorted Inception feature variance per attribute. Middle: Per-attribute mean IvOM over 30,000 CelebA training samples for StyleGAN2 (red) and for our method (blue), where each bar corresponds to one attribute.

we first show the existence of biases across CelebA attributes in terms of sample counts and sample variance, and then correlate them to the biased performance of the backbone StyleGAN2 [23].

As shown in the left barplot of Figure 2, given the attribute histogram over 30,000 samples, 29 out of 40 binary attributes are more than 50% biased from the balance point (15,000 out of 30,000 samples with a positive attribute annotation, shown as the red dashed line). On the other hand, in the right barplot of Figure 2, we calculate the standard deviation of Inception features [40] of samples within each attribute, and notice a wide range spanning from 0.038 to 0.062.

Too few samples or too large appearance variance in one attribute discourages generative coverage for that attribute, and thus results in biases. To quantify the per-attribute coverage, we measure the mean IvOM [33] over positive training samples. A larger value indicates a worse coverage. In the middle barplot of Figure 2, we visualize the correlation between IvOM and the joint distribution of sample counts and sample variance. There is a clear gradient trend of IvOM when the samples of an attribute turn rarer and/or more diverse. To validate such a strong correlation, we first normalize the sample counts and sample variance across attributes by their means and standard deviations. Then we simply add them up as a joint variable vector, and calculate its Spearman's ranking correlation to the per-attribute IvOM. For StyleGAN2 (the red bar), the correlation coefficient of 0.75 indicates a strong correlation between data biases and model biases. This evidences the urgency to mitigate biases against the rare and diverse samples, in another word, to enhance the coverage over minority subgroups.

Table 2. Comparisons on CelebA dataset. We indicate for each metric whether a higher (⇑) or lower (⇓) value is more desirable. The first part corresponds to the comparisons among different methods. For VAEGAN we report the results based on LPIPS distance metric. We report additional results based on the other three metrics in the supplementary material. We highlight the best performance in **bold** and the second best performance with underline. We visualize the radar plots in Figure 3 for the comprehensive evaluation of each method over the validation set. The second part corresponds to our minority inclusion model variants in Section 4.5.

	FID30k		Precision30k		Recall30k		IvOM3k		IvOM3k std	
	⇓		⇑		⇑		⇓		⇓	
Method	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val
StyleGAN2 [23]	**9.37**	**9.49**	0.855	0.844	0.730	0.741	0.303	0.302	0.0268	0.0264
SNGAN [34]	13.32	13.24	0.792	0.787	0.631	0.616	0.325	0.322	0.0274	0.0261
Dist-GAN [43]	30.97	30.44	0.511	0.595	0.360	0.385	0.282	0.280	0.0220	0.0209
DSGAN [46]	14.29	14.00	0.868	0.862	0.679	0.724	0.301	0.300	0.0227	0.0220
PacGAN [30]	15.05	15.12	0.870	0.869	0.726	0.758	0.311	0.308	0.0256	0.0238
ALI [12]	10.09	10.06	0.842	0.867	0.688	0.710	0.298	0.297	0.0240	0.0245
VAEGAN [26] LPIPS	24.10	23.47	0.878	0.851	0.572	0.560	0.318	0.315	0.0284	0.0272
α-GAN [37]	12.65	12.53	0.803	0.810	0.757	0.763	0.267	0.267	0.0208	0.0192
VEEGAN [42]	16.34	16.13	0.752	0.768	0.660	0.695	0.260	0.269	**0.0190**	**0.0181**
Ours LPIPS interp	11.56	11.28	**0.927**	**0.941**	**0.849**	**0.848**	**0.255**	**0.262**	0.0193	0.0195
Ours *Eyeglasses*	13.54	14.43	0.914	0.910	0.890	0.895	0.255	0.265	0.0249	0.0193
Ours *Bald*	13.34	13.46	0.903	0.895	0.886	0.892	0.268	0.272	0.0381	0.0227
Ours *EN&HM*	15.18	15.00	0.885	0.891	0.830	0.842	0.268	0.270	0.0318	0.0277
Ours *BUE&HC&A*	14.27	13.85	0.878	0.874	0.871	0.884	0.262	0.266	0.0300	0.0254

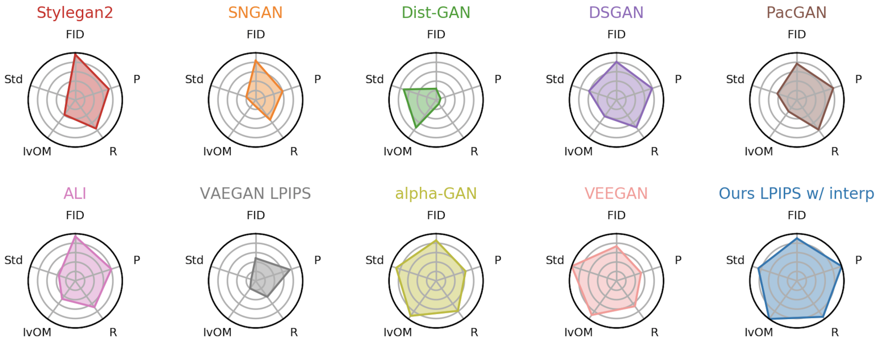

Fig. 3. Radar plots for the first part of Table 2. "P" represents Precision, "R" represents Recall, and "Std" represents IvOM standard deviation. Values have been normalized to the unit range, and axes are inverted so that the higher value is always better.

4.4 Comparisons on CelebA

In Section 3.3 we propose two strategies to harmonize adversarial and reconstructive training: the deep distance metric and the interpolation-based aug-

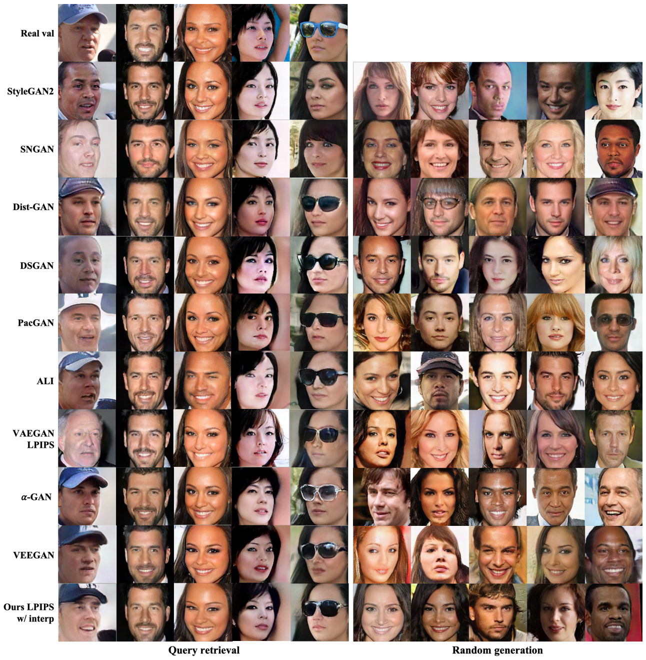

Fig. 4. Retrieval samples on the left (used for IvOM evaluation) and random generation samples on the right (used for FID, precision, and recall evaluation). The query images for retrieval in the top left row are real and unseen during training.

mentation. We compare four distance metrics and with/without augmentation in the supplementary material. We obtain: (1) LPIPS similarity shows near-top performance all around measures; and (2) interpolation-based augmentation consistently benefits all the measures in general for all the distance metrics. We therefore employ both into our full method.

To evaluate our data coverage performance in practice, we conduct comprehensive comparisons on CelebA [31] against baseline methods. The first part of Table 2 show our comparisons. Figure 3 assists interpret the table. We find:

(1) FID is not a gold standard to reflect the entire capability of a generative model, as it ranks differently from the other metrics.

(2) Compared to the original backbone StyleGAN2 which achieves the second-best FID, our full method ("Ours LPIPS interp") trades slight FID deterioration for significant boosts in all the other metrics. This is meaningful because precision (FID) can be traded off at the expense of recall (Recall, IvOM) via the truncation trick used in [6,23], while the opposite direction is infeasible.

(3) Our full method outperforms all the existing state-of-the-art techniques in terms of Precision, Recall, and IvOM, where the latter two are the key evidence for effective data coverage. The last radar plot in Figure 3 shows our method achieves near-top measures all around with the most balanced performance.

Table 3. Comparisons on CelebA minority subgroups, where the percentages show their portion w.r.t. the entire population. The metrics are measured on the corresponding subgroups only. We indicate for each metric whether a higher (⇑) or lower (⇓) value is more desirable. We highlight the best performance in **bold**.

Arbitrary minority subgroup	Method	Precision1k minority only ⇑ Train	Val	Recall1k minority only ⇑ Train	Val	IvOM1k minority only ⇓ Train	Val
	StyleGAN2 [23]	0.719	0.704	0.582	0.589	0.355	0.352
Eyeglasses	Ours LPIPS interp	0.843	0.845	0.740	0.708	0.309	0.308
(6%)	Ours *Eyeglasses*	**0.904**	**0.919**	**0.897**	**0.892**	**0.261**	**0.288**
	StyleGAN2 [23]	0.707	0.750	0.461	0.424	0.301	0.305
Bald	Ours LPIPS interp	0.763	**0.783**	0.666	0.670	0.269	**0.273**
(2%)	Ours *Bald*	**0.779**	0.718	**0.842**	**0.810**	**0.189**	**0.273**
Narrow_Eyes	StyleGAN2 [23]	0.719	0.701	0.543	0.577	0.272	0.274
&*Heavy_Makeup*	Ours LPIPS interp	0.794	0.760	0.632	0.621	0.246	0.248
(4%)	Ours *EN&HM*	**0.799**	**0.766**	**0.698**	**0.696**	**0.194**	**0.244**
Bags_Under_Eyes	StyleGAN2 [23]	0.838	0.804	0.736	0.725	0.263	0.268
&*High_Cheekbones*	Ours LPIPS interp	0.816	0.831	0.700	0.742	0.237	0.241
&*Attractive* (4%)	Ours *BUE&HC&A*	**0.889**	**0.883**	**0.813**	**0.809**	**0.191**	**0.237**

(4) Our method also achieves the top-3 performance in the standard deviation of per-attribute IvOM, indicating an equalized capacity across the attribute spectrum. The blue bars in the middle barplot of Figure 2 also visualize our method consistently outperforms StyleGAN2 (red bars) for all the attributes, in particular with more significant improvement for the minority subgroups.

(5) Figure 4 shows qualitative comparisons in terms of query retrieval and uncurated random generation. StyleGAN2 suffers from mode collapse. For the collapsed modes, our method significantly improves the generation from non-existence of rare attributes to good quality (hat, sunglasses, etc.). Our method also demonstrates desirable generation fidelity and diversity.

(6) All the conclusions above generalize well to unseen data, as evidenced by the "Val" columns in Table 2.

4.5 Extension to Minority Inclusion

We adapt our method for ensuring specific coverage over minority subgroups (Algorithm 1). Without introducing unconscious bias on the CelebA attributes, we arbitrarily specify four sets of attributes, the samples of which count for no more than 6% of the population, and therefore, constitute four minority subgroups respectively. The attribute sets and their portions are listed in the first column of Table 3.

To validate minority inclusion, we first compare our minority model variants over the corresponding minority subsets against the backbone StyleGAN2 and

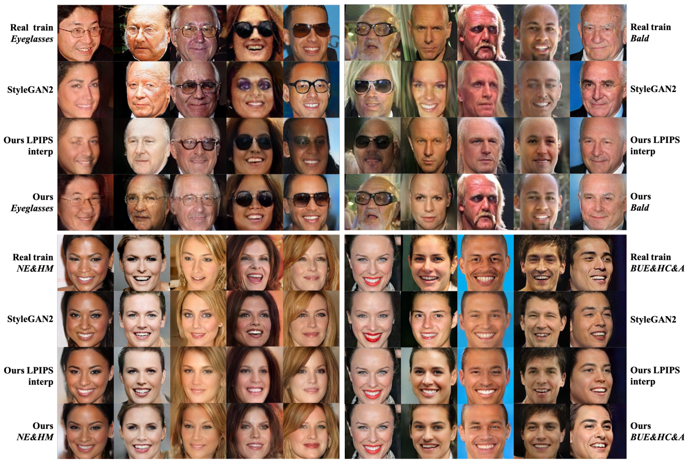

Fig. 5. Retrieval samples according to different minority subgroups. The query images for retrieval in the top row of each sub-figure are real from the training set.

against our general full model. See Table 3 for the results. Our minority variants consistently outperform the two baselines over all the minority subgroups. In Figure 5, our method retrieves the minority attributes the most accurately, even for the subtle attributes like eye bags where StyleGAN2 fails. It validates better training data utilization of our minority models. Additional results are shown in the supplementary material and supplementary videos.

To validate the overall performance beyond minority subgroups, we show at the bottom of Table 2 the performance on the entire attribute spectrum. We conclude that the improvement of all our minority models comes at little or no compromise from their performance on the overall dataset.

5 Conclusion

In this paper, we formalized the problem of minority inclusion as one of data coverage and improved data coverage using a novel paradigm that harmonizes adversarial training (GAN) with reconstructive generation (IMLE). Our method outperforms state-of-the-art methods in terms of Precision, Recall, and IvOM on CelebA, and the improvement generalizes well on unseen data. We further extended our method to ensure explicit inclusion for minority subgroups at little or no compromise on overall full-dataset performance. We believe this is an important step towards fairness in generative models, with the aim to reduce and ultimately prevent discrimination due to model and data biases.

Acknowledgement. This project was partially funded by DARPA MediFor program under cooperative agreement FA87501620191 and by ONR MURI (N00014-14-1-0671). We thank Tero Karras and Michal Lukáč for sharing code. We also thank Richard Zhang and Dingfan Chen for constructive advice in general.

References

1. Adler, J., Lunz, S.: Banach wasserstein gan. In: NeurIPS (2018) 4
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. In: ICML (2017) 4
3. Berthelot, D., Schumm, T., Metz, L.: Began: Boundary equilibrium generative adversarial networks (2017) 4
4. Bhattacharyya, A., Fritz, M., Schiele, B.: "best-of-many-samples" distribution matching. arXiv (2019) 2
5. Bowman, S., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. In: SIGNLL (2016) 3
6. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: ICLR (2019) 12
7. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: ICDM Workshops (2009) 3
8. Chen, L., Dai, S., Pu, Y., Zhou, E., Li, C., Su, Q., Chen, C., Carin, L.: Symmetric variational autoencoder and connections to adversarial learning. In: AISTATS (2018) 4
9. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: NeurIPS (2016) 4
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 1
11. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. In: ICLR (2016) 1, 4
12. Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., Courville, A.: Adversarially learned inference. In: ICLR (2016) 1, 4, 8, 9, 11
13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference (2012) 3
14. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: KDD (2015) 3
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014) 1, 4
16. Grover, A., Choi, K., Shu, R., Ermon, S.: Fair generative modeling via weak supervision. arXiv (2019) 1, 3
17. Grover, A., Song, J., Kapoor, A., Tran, K., Agarwal, A., Horvitz, E.J., Ermon, S.: Bias correction of learned generative models using likelihood-free importance weighting. In: NeurIPS (2019) 1, 3
18. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: NeurIPS (2017) 4
19. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: NeurIPS (2016) 3
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017) 9
21. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: ICDM (2010) 3
22. Kamishima, T., Akaho, S., Sakuma, J.: Fairness-aware learning through regularization approach. In: ICDM Workshops (2011) 3

23. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. arXiv (2019) 2, 8, 9, 10, 11, 12, 13
24. Kim, Y., Wiseman, S., Miller, A.C., Sontag, D., Rush, A.M.: Semi-amortized variational autoencoders. In: ICML (2018) 3
25. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014) 1, 4
26. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: ICML (2016) 1, 2, 4, 5, 7, 8, 9, 11
27. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE (1998) 8
28. Li, K., Malik, J.: Fast k-nearest neighbour search via prioritized dci. In: ICML (2017) 5
29. Li, K., Malik, J.: Implicit maximum likelihood estimation. arXiv (2018) 1, 2, 4, 5
30. Lin, Z., Khetan, A., Fanti, G., Oh, S.: Pacgan: The power of two samples in generative adversarial networks. In: NeurIPS (2018) 4, 8, 9, 11
31. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015) 1, 8, 12
32. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning (2019) 3
33. Metz, L., Poole, B., Pfau, D., Sohl-Dickstein, J.: Unrolled generative adversarial networks. In: ICLR (2017) 4, 8, 9, 10
34. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR (2018) 4, 8, 9, 11
35. Peng, X.B., Kanazawa, A., Toyer, S., Abbeel, P., Levine, S.: Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow (2019) 4
36. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and variational inference in deep latent gaussian models. In: ICML (2014) 1, 4
37. Rosca, M., Lakshminarayanan, B., Warde-Farley, D., Mohamed, S.: Variational approaches for auto-encoding generative adversarial networks. arXiv (2017) 2, 4, 8, 11
38. Ryu, H.J., Adam, H., Mitchell, M.: Inclusivefacenet: Improving face attribute detection with race and gender diversity (2018) 1, 3
39. Sajjadi, M.S., Bachem, O., Lucic, M., Bousquet, O., Gelly, S.: Assessing generative models via precision and recall. In: NeurIPS (2018) 9
40. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NeurIPS (2016) 7, 10
41. Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., Varshney, K.R.: Fairness gan: Generating datasets with fairness properties using a generative adversarial network (2019) 1, 3
42. Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U., Sutton, C.: Veegan: Reducing mode collapse in gans using implicit variational learning. In: NeurIPS (2017) 2, 4, 5, 8, 9, 11
43. Tran, N.T., Bui, T.A., Cheung, N.M.: Dist-gan: An improved gan using distance constraints. In: ECCV (2018) 4, 8, 11
44. Warde-Farley, D., Bengio, Y.: Improving generative adversarial networks with denoising feature matching. In: ICLR (2017) 4
45. Xu, D., Yuan, S., Zhang, L., Wu, X.: Fairgan: Fairness-aware generative adversarial networks. In: Big Data (2018) 1, 3
46. Yang, D., Hong, S., Jang, Y., Zhao, T., Lee, H.: Diversity-sensitive conditional generative adversarial networks. In: ICLR (2019) 4, 8, 9, 11

47. Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv (2015) 1
48. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: AISTATS (2017) 3
49. Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI (2017) 3
50. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 7
51. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: EMNLP (2017) 3
52. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network. In: ICLR (2017) 4
53. Zhao, S., Ren, H., Yuan, A., Song, J., Goodman, N., Ermon, S.: Bias and generalization in deep generative models: An empirical study. In: NeurIPS (2018) 1, 3