Debiasing Image-to-Image Translation Models

Md Mehrab Tanjim[1]
mtanjim@eng.ucsd.edu
Krishna Kumar Singh[2]
krishsin@adobe.com
Kushal Kafle[2]
kkafle@adobe.com
Ritwik Sinha[2]
risinha@adobe.com
Garrison W. Cottrell[1]
gary@eng.ucsd.edu

[1] University of California San Diego
9500 Gilman Dr,
La Jolla, CA, USA

[2] Adobe Research
345 Park Ave,
San Jose, CA, USA

Abstract

Deep generative models have shown a lot of promise in various image-to-image translation tasks such as image enhancement and generating images from sketches. However, when all the classes are not equally represented in the training data, these algorithms can fail for underrepresented classes. For example, our experiments with the CelebA-HQ face dataset reveal that this bias is prevalent for infrequent attributes, e.g., eyeglasses and baldness. Even when the input image clearly has eyeglasses, the image translation model is unable to create a face with them. To remedy this problem, we propose a data and model agnostic, general framework based on contrastive learning, re-sampling, and minority category supervision to debias existing image translation networks for various image-to-image translation tasks such as super-resolution and sketch-to-image. Our experimental results from the real and synthetic datasets show that our framework outperforms the baselines both quantitatively and qualitatively.

1 Introduction

Generative Adversarial Networks (GANs) [11] have shown significant promise in synthesizing high-fidelity images [20, 21]. As a result, they have been adapted to achieve stunning results in many image-to-image translation (I2I) tasks, such as super-resolution [24, 39, 44], sketch-to-image [6, 8, 30], image inpainting [10, 45], etc. In these tasks, most current work focuses on the quality of generated results.

In this work, we study the capacity of existing image-to-image translation models to generate attributes that are in the minority in the training set. Figure 1 shows examples from the Pixel2Style2Pixel (pSp) model [29] network, which is one of the most popular and successful I2I models. The results for super-resolution and sketch-to-image tasks show an incredible visual quality of synthesized images, but also show an utter failure to generate

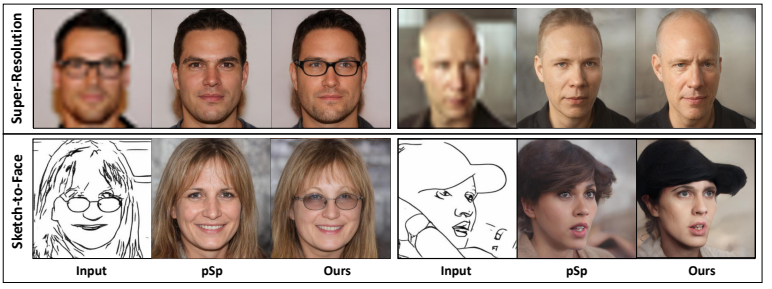


Figure 1: Examples of how Pixel2Style2Pixel (pSp) [29] is biased against minority attributes in the CelebA-HQ dataset [19]. We also show results from our debiasing framework (Ours).

minority visual attributes, such as eyeglasses (about 5% of the data) or baldness (about 2%), despite being clearly visible in the low-resolution or sketch input.

We have found that this problem is not limited to one particular architecture or dataset; whenever there is class imbalance in the training set, existing I2I translation models exhibit similar limitations. For example, we have trained the popular I2I translation model, pix2pix [17], on two synthetic datasets and found that the generated images are biased towards majority attributes. This bias in I2I translation models can have a negative impact on various important downstream applications (e.g., image enhancement by super-resolution).

Here, we identify the need to debias I2I translation models, and propose a general framework to solve it. Normally used debiasing techniques, such as re-sampling [4, 36] and auxiliary classifier loss [27, 32], have not been explored for I2I. In this paper, we showcase their success for I2I tasks. However, note that these methods only operate at the input level (re-sampling), or at final generation (pre-trained classifier). Without additional constraint in between, the latent features of biased classes can still become similar to the codes of the non-biased classes at the encoding level. This can prevent the decoder from learning a proper mapping between the latent codes and the output images from minority and majority classes during the generation process. To overcome this issue, we apply supervised contrastive learning during training to separate the encoded features of the minority from the majority, which helps the decoder to capture the features necessary to generate images with the correct attributes for both classes. We further conduct extensive experiments to show the effectiveness of our method on super-resolution and sketch-to-image I2I tasks. Figure 1 also shows how our method overcomes the bias problem for the pSp network. Note that our framework is agnostic to the particular encoder/decoder architecture.

Contributions: 1) We identify the bias problem in image-to-image translation and propose a new task of debiasing these models. 2) We propose a novel contrastive learning-based approach which outperforms the baselines both quantitatively and qualitatively. 3) Finally, we show that our model generalizes well, we apply it to multiple image-to-image translation tasks and datasets.

2 Related Work

Image-to-Image Translation. The goal of image-to-image translation models is to map images from a source domain (e.g., low resolution input) to images of a target domain (e.g., high resolution output), i.e., conditional image generation. The most notable work in this direction is pix2pix [17], where they show that conditional GANs can be used to solve a wide

variety of image-to-image translation tasks. Motivated by their success, researchers have adopted specific conditional GAN architectures to solve specific image-to-image translation problems, for example, super-resolution [24, 39, 44], sketch-to-image [5, 6, 30], semantic label-to-image [26, 28, 49], unpaired translation [25, 48], or multi-modal image synthesis [7, 15]. However, most of these models are application-specific and may not generate high-resolution outputs.

For this problem, Pixel2Style2Pixel (pSp) [29] achieves promising results. Motivated by the capabilities of StyleGAN2 [21] to generate photo-realistic images, they train an encoder to project source images into the latent space of a pre-trained StyleGAN2 to solve image-to-image translation tasks. They show the effectiveness of their approach on tasks such as super-resolution, sketch-to-image, face frontalization, etc. Hence, we have chosen their model for conducting our experiments. Additionally, we have used popular pix2pix model [17] to test the generalizability of our approach.

Debiasing Frameworks. Bias and fairness have recently received a great deal of attention in the research community. Researchers have identified how the training datasets can suffer from various biases [22, 37, 38] and how it can lead to undesired behaviors in various image classification networks [1, 9, 13, 33, 34]. Some recent work explores using generative models to create more balanced datasets for recognition tasks. For example, [3] creates synthetic images with latent space exploration so that bias in the classification network can be algorithmically measured. Similarly, to mitigate bias in classification networks, [2] adapted a variational autoencoder to learn the probability distribution of latent features. Based on the probability distribution, they re-sample those latent images which have lower probability to balance the dataset during training. There is limited work that aims to reduce bias for image generation itself; mostly, work has focused on the unconditional generation task to generate less biased distributions [31, 46]. All of these approaches focus on creating a balanced dataset, rectifying problems in the classification network, or doing unconditional generation.

In the case of debiasing image translation models, very few frameworks exist. [18] proposed debiasing I2IT models by using posterior sampling via gradient optimization, i.e., finding the optimal latent codes given the input. However, their application is limited in the case of reconstruction from noisy input, such as denoising or super-resolution. Furthermore, their debiasing method is not applicable to encoder-decoder architectures, which is the common architecture for I2IT models. Similarly, other works [16, 40] are either specific to model architectures, thus limiting the scope to apply their ideas to latest SOTA I2IT models, or need the generator to be retrained for learning a debiased representation. In this work, we propose a framework that can be applied to any I2IT tasks and models. Our proposed framework intercepts the encoding stage and it can even work for frozen generators (e.g., a frozen StyleGAN2 generator). Additionally, unlike all previous works, our framework can debias while maintaining high-quality. We describe our proposed framework in the following section.

3 Approach

In order to tackle the bias problem in the image-to-image translation models, we follow the common setting for the most of the debiasing work (e.g. [1, 9, 13, 33, 34]): we assume the bias is known or can be measured. In this section, we first discuss how we measure bias in the Celeb-HQ dataset for the image-to-image translation task. Next, we introduce our debiasing framework.

Attribute	Bald	Wearing Hat	Eyeglasses	Blond Hair	Bangs	Black Hair	Male	Heavy Makeup	High Cheekbones	Smiling
Percentage ↓	**2.37**	**3.57**	**4.89**	17.09	18.08	21.97	36.86	45.69	46.16	46.97
F1 Score on Real	0.8142	0.8908	0.9825	0.8483	0.8756	0.8186	0.9791	0.8906	0.8576	0.9333
F1 Score on Generated ↓	**0.7216**	**0.2500**	**0.0984**	0.8288	0.8393	0.7725	0.9653	0.8444	0.8235	0.9089

Table 1: Bias analysis in the CelebA-HQ [19] dataset. The least three values in *Percentage* and *F1 Score on Generated* are shown in bold.

3.1 Measuring the Bias

We use two main criteria to detect bias. The first criteria is simple and straightforward: sort the attributes by the fraction of images in which they occur and select the attributes with lower fractions. Naturally, if images for a particular attribute are rare, then generation can become skewed against it. The second criteria is based on the ability to detect the attribute. This is important because we would like to reliably evaluate the performance of the model, and thus, we should select the attributes for which the classifiers show high accuracy.

To apply these criteria, we first train a ResNet152 classifier [12] on the same training set as Pixel2Style2Pixel [29] (pSp) and filter any attribute that shows a low F1 score (using a threshold of 0.8). For example, we observed an F1 classifier score higher than 0.95 for 'Eyeglasses' but a score of 0.0 for 'Blurry' which indicates 'Eyeglasses' clearly has better image feature representation than 'Blurry.' We therefore choose attributes that have high F1 scores and are rare in the dataset, providing us with rare attributes that are easily labeled automatically. To validate that being a minority in the training dataset can pose a bias problem for I2I task, we calculate the classifier F1 scores on generated images and assess the performance on minority classes. Table 1 shows F1 scores on generated images along with the percentage of biased class and classifier F1 scores on the ground truth. Unsurprisingly, the under-performance is most pronounced in the attributes that are rare. There is a significant drop in the F1 scores between the real and generated images, when the attribute is in less than 5% of the images. We also see this qualitatively in Figure 1. Additional examples are shown in the supplementary material. Here, we can see that none of the generated images from the pre-trained pSp network for either of the tasks faithfully reconstruct the attributes, although they are clearly visible in the input images. We also run the classifier on the input images to verify this (scores are included in the supplementary). This indicates the input images have information about the attributes and the pSp network should be able to reconstruct them. In the presence of bias, the model can also undesirably add some features. We can see from Figure 1, this is the case for 'Bald', where the model is hallucinating hair. These errors can have a detrimental impact on various downstream applications. Therefore, in this work, we address this problem and propose a general debiasing framework. For debiasing, we select the first 3 attributes from Table 1, i.e.,'Bald', 'Wearing Hat', and 'Eyeglasses.' It is worth mentioning, although different tasks might require different ways of measuring bias, our debiasing framework is designed to be agnostic of how the bias is measured. In the following, we discuss our proposed debiasing framework.

3.2 Our Debiasing Framework

To mitigate bias in existing image-to-image translation models, the first insight comes from the exploration of images in the latent space. When the images from a certain group (e.g., images with eyeglasses or bald) are rare, their representation in the latent space can become similar to the majority group. As a result, the model becomes biased to frequent patterns. To remedy this problem, the key idea is to separate the latent codes for minority and majority

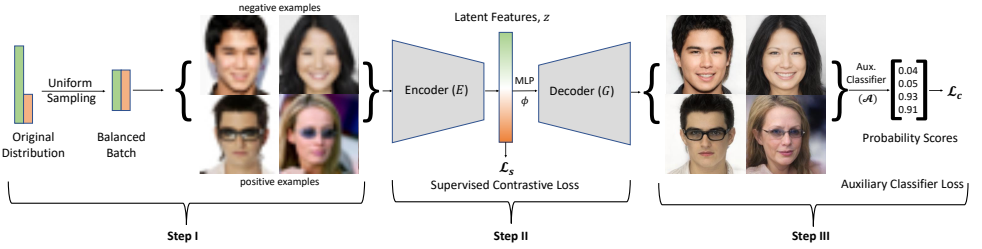


Figure 2: Our proposed debiasing framework. We first start by creating a balanced batch for a given attribute/class (Step I). Then, we apply supervised contrastive loss on the latent features (Step II). Finally, we apply an auxiliary classifier loss based on the prediction of the attribute/class on the generated images (Step III).

groups, allowing the model to generate from their respective distributions independently.

A simple way to solve this during training is to over-sample the minority. Over-sampling the minority forces the network to see more instances of rare images and helps it encode the attributes. This is our first step towards debiasing is Step I in Figure 2. Although re-sampling is considered a general-trick for class imbalance problems, its effectiveness in image translation tasks has not been explored before. In some cases, it has been shown to not be effective [35, 36]. To improve on this, we propose using metric learning based losses [8, 42] to further separate the latent codes from different groups or classes. Here, we apply supervised contrastive loss [23]. This loss pulls together the representation of images from the same class (whether minority or majority) in the latent space and pushes them apart if from different classes. Mathematically,

$$\mathcal{L}_s = -\sum_{i\in\mathcal{I}} \frac{1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(z_i \cdot z_p)/\tau}{\sum_{x\in\mathcal{X}(i)} \exp(z_i \cdot z_x)/\tau} \tag{1}$$

Here, $i \in \mathcal{I} \equiv 1,\ldots,N$ (N is the batch size) is the index of an arbitrary sampled image I_i from the set of all images $\mathcal{I}$, $\mathcal{X}(i) \equiv \mathcal{I}/\{i\}$, $P(i) \equiv \{p \in \mathcal{X}(i) : y_p = y_i\}$ (y is the binary label or class of that image), $z_i = E(I_i)$ is the latent feature representation of the i^{th} image, I_i after it goes through the encoder E, and τ is a scalar temperature parameter. The positive pairs z_p in the supervised contrastive loss are obtained from the images that belong to same class and negative pairs are the images that belong to different classes. For example, images for 'Eyeglasses' will have positive pairs among themselves and images without 'Eyeglasses' will be the negative pairs in this case (shown as positive and negative examples in Figure 2). We should mention that we L_2 normalize the latent features to get the corresponding directions for applying supervised contrastive loss. However, the unit vectors or directions may not be suitable for generation. For this purpose, we pass the latent codes, z_i, through a multi-layer perceptron (MLP) layer, ϕ, after we have applied the $\mathcal{L}_s$ loss. The decoder or generator G then takes $\phi(z_i)$ as inputs for generating the target images. This is the second step in our framework (Figure 2 Step II).

Although we apply the contrastive loss in a supervised manner, and we are separating the latent codes of minorities and majorities, this may not always give the generator, G, enough incentive to focus on generating the particular attribute. So, to enforce this constraint further during the generation process, we use an auxiliary classifier $\mathcal{A}$ to predict the desired attribute

on the generated images, $G(\phi(z_i))$, from the decoder and apply binary cross entropy loss:

$$\mathcal{L}_c = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y_i \cdot \log(\hat{y}_i) + (1-y_i)\cdot(1-\log(\hat{y}_i)), \tag{2}$$

where $\hat{y}_i = \mathcal{A}(G(\phi(z_i)))$. This can further assist the supervised contrastive loss, $\mathcal{L}_s$, to separate the latent codes such that the desired attribute can be generated more easily. The final loss function is as follows:

$$\mathcal{L} = \mathcal{L}_o + \lambda_s * \mathcal{L}_s + \lambda_c * \mathcal{L}_c, \tag{3}$$

where $\mathcal{L}_o$ is the original loss function used to train the image-to-image translation model without our changes, $\mathcal{L}_s$ is the supervised contrastive loss, and $\mathcal{L}_c$ is the auxiliary binary cross entropy loss. The hyperparameters λ_s and λ_c balance the different losses. One thing to note from our debiasing steps is that our framework has no dependency on a specific encoder-decoder architecture. Thus, our approach generalizes to any image translation model.

4 Experiments

Dataset. We experiment on datasets where the bias occurs naturally. We also create two synthetically biased datasets.

1) CelebA-HQ. For experiments with human faces, we have selected the CelebA-HQ [19] dataset. As mentioned previously, in this dataset, the bias occurs naturally. The details on our train-validation-test split are described in the supplementary. *2) Bags and Shoes.* Our first synthetic dataset consists of images of bags from 'edge2bags' [17] and shoes from 'edge2shoes' [17]. We have selected a total of 5000 images, where 4950 images belong to 'Shoes' category, and the remaining 50 are from 'Bags' (99:1 bias ratio). We call this dataset 'Bags and Shoes.' We separately keep 200 images in total for both validation and test set. *3) Cats and Dogs.* For this dataset, we select animal faces from AFHQ [7]. Specifically, we have selected faces of Cats and Dogs. The training, validation, and testing split follow the same strategy as 'Bags and Shoes.' In this dataset, the majority class is 'Cats.'

Tasks and Models. For experiments with faces, we select two popular image-to-image translation tasks, namely, super-resolution and sketch-to-face. For performing these translation tasks on human faces, we have selected one of the recently proposed image-to-image translation models, Pixel2Style2Pixel (pSp) [29]. As debiasing the image-to-image translation task is new, there are no existing baselines. Hence, for comparison, we created our own baseline and variants of our model: **1) Vanilla.** Original pSp network without any changes. We train the network from scratch on our dataset for each of the attributes. **2) Sampling Baseline.** This is a trivial baseline where images from the minority class are resampled to create a balanced batch (our Step I). We also apply data augmentations (e.g. shifting, shearing, scaling, horizontal flipping, etc.) to all images when re-sampling. **3) Ours (I+II)** In this model, we take the first two steps from our pipeline, that is re-sampling and applying supervised contrastive loss (Equation 1), $\mathcal{L}_s$, during the encoding-decoding phase. **4) Ours (I+III)** Here, we only consider re-sampling and applying auxiliary loss (Equation 2), $\mathcal{L}_c$. **5) Ours (I+II+III)** All three components in our debiasing framework.

For all of our experiments with pSp, we use the same frozen StyleGAN2 as the decoder, pre-trained on FFHQ (which has good coverage for accessories like eyeglasses, hats, etc.). Using this network, we were able to generate rare features in Celeb-HQ such as eyeglasses,

hats, etc. (Figure 1, 3). This shows that the latent codes are already available in the pre-trained StyleGAN2 generator network. Any problems, therefore, lie in pSp's encoder, which becomes biased during training with a biased dataset (e.g. Celeb-HQ).

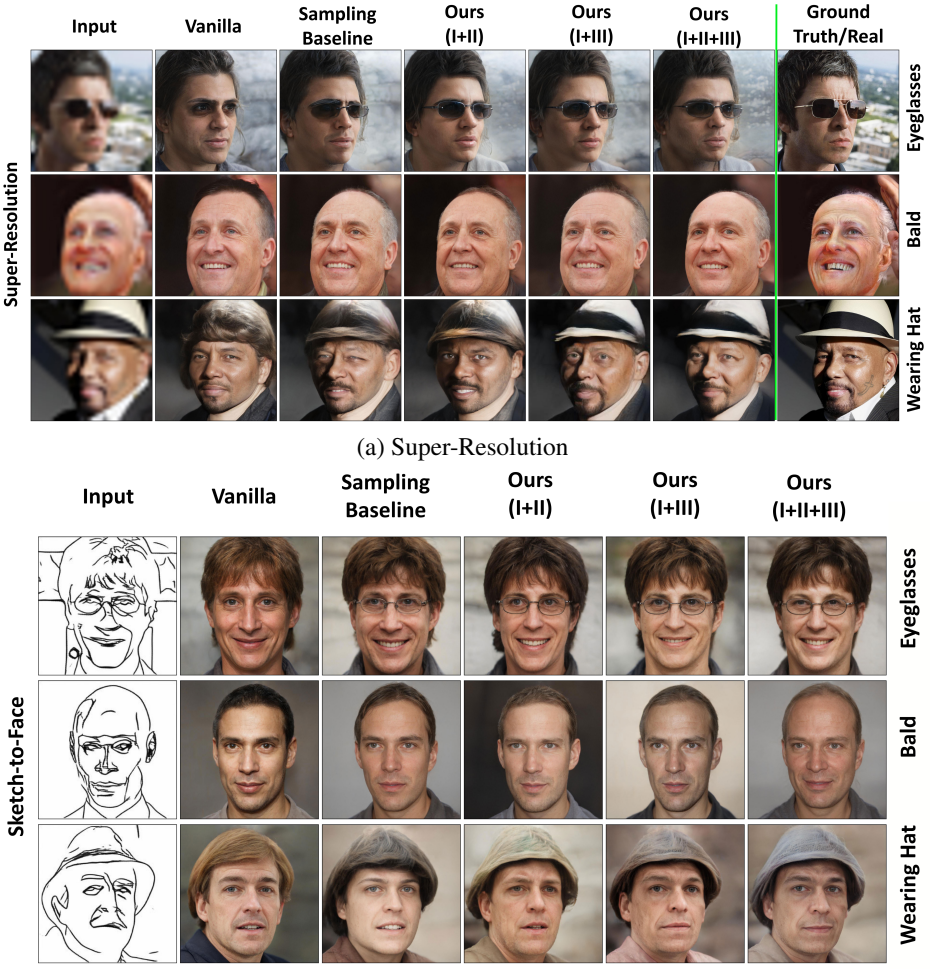
Our debiasing framework is not only limited to the pSp network and human faces. We can apply it to different I2I translation architectures, and images from domains other than human faces. To show this, we have chosen another popular image-to-image translation network, pix2pix [17], and perform edge-to-image task on our synthetic datasets, namely 'Bags and Shoes', and 'Cats and Dogs.' Similar to pSp, we create different variants of our model for pix2pix. More details of our data augmentations and training procedures for both models can be found in the supplementary.

Evaluation. To measure how well the models faithfully generate the attribute (or absence of it), we report the classifier prediction scores on the generated images. We convert all scores to the same scale (between 0 and 100). A model is better if it obtains high scores on the minority group while maintaining the majority group performance. For fairness, we keep the classifiers for evaluation different from our models (more details in the supplementary). For the super-resolution task with pSp, we also report Learned Perceptual Image Patch Similarity (LPIPS) [47] and MSE in order to evaluate whether our generated images overall match the target images. For experiments with pix2pix, we perform edge-to-image synthesis; given the loss of information in this task, we do not consider there to be a 'ground truth' image, so LPIPS/MSE do not apply. In this case, we report Fréchet Inception Distance (FID) [14] to measure if the generated images match the actual distribution of their respective classes.

4.1 Quantitative and Qualitative Results

Table 2 shows the results. Our model and its variations always outperform the Vanilla and Sampling baselines for the minority groups. In terms of majority performance, Vanilla's performance is often better, as one would expect since it is biased to the majority class. The table also reveals the individual contribution of all three components of our model. For example, applying only supervised contrastive loss on top of re-sampling (**I+II**) helps in almost all cases. Applying the auxiliary classifier loss (**I+III**) helps even more in 11 out of 18 cases. But when supervised contrastive loss is applied with the auxiliary classifier (**I+II+III**), it improves the minority class scores for almost all cases, with a negligible difference in the remaining case. For example, it leads to about 30% and 57% improvement in super-resolution and sketch-to-face, respectively, for the 'Wearing Hat' minority class compared to the sampling baseline. In this new task, general tricks like re-sampling and auxiliary classifier help a lot, which is not always the case in many debiasing tasks [35, 36, 41, 43], but adding the contrastive loss generally improves on these.

For tasks like super-resolution, it is also important to match the quality of generated images with the ground truth. Therefore, we report LPIPS [47] and MSE in Table 3. We can see our framework performs debiasing without compromising the image quality. For further qualitative comparison, we show the generated images from each of the models for each attribute and task in Figure 3. We also show the ground truth images for super-resolution in the right most column of 3 (a). We can see, in all cases, the Vanilla model does not produce the desired outputs. Compared to other alternatives, we can see our methods produce much better results. For example, 'Wearing Hat' appears to be the hardest attribute to reconstruct, for both tasks. Even for this attribute, we can see our contrastive model (I+II+III) is producing hat-like shapes and textures. Additionally, we can see the quality of the images from our model is similar or better than Vanilla in all cases, which suggests that our framework is an



(a) Super-Resolution

(b) Sketch-to-Face

Figure 3: Results of our debiasing framework compared to the Vanilla and Sampling Baseline model. Here, we show one example for each of the considered tasks across all attributes. Our generated results better capture the attributes compared to baselines.

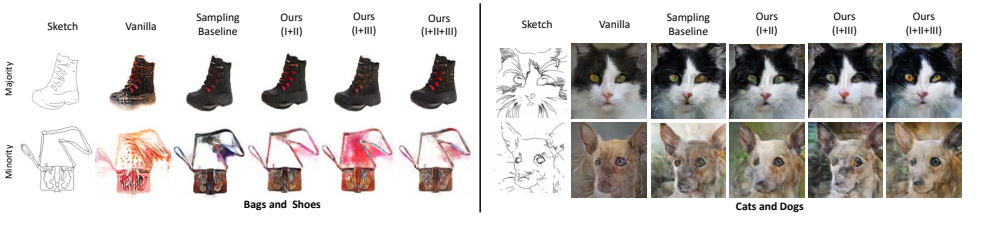


Figure 4: Our debiasing framework is not only limited to a particular model. Here, we show how our idea can be applied to pix2pix [] to improve the quality of synthetic images in the presence of bias.

Task	Attribute	Group	Vanilla	Sampling Baseline	Ours (I+II)	Ours (I+III)	Ours (I+II+III)
Super-Resolution	Eyeglasses	Minority	15.27	89.95	91.55	92.35	**92.85**
		Majority	98.52	98.21	98.6	97.8	**98.7**
		Both	56.89	94.08	95.08	95.08	**95.77**
	Bald	Minority	63.91	88.89	90.41	90.12	**91.60**
		Majority	**98.18**	97.53	98.01	98.01	98.01
		Both	81.04	93.21	94.21	94.06	**94.81**
	Wearing Hat	Minority	23.19	60.71	61.95	78.31	**80.84**
		Majority	**98.4**	97.62	97.93	97.95	98.21
		Both	60.8	79.17	79.94	88.13	**89.52**
Sketch-to-Face	Eyeglasses	Minority	30.32	92.73	93.30	**94.10**	94.06
		Majority	98.65	96.05	97.99	98.7	**98.84**
		Both	64.48	94.39	95.64	96.39	**96.45**
	Bald	Minority	46.88	85.26	81.03	86.80	**89.12**
		Majority	**98.3**	95.47	97.09	96.46	95.53
		Both	72.59.	90.36	89.06	91.63	**92.32**
	Wearing Hat	Minority	15.58	32.38	50.49	78.61	**79.50**
		Majority	**98.37**	98.07	97.69	97.01	97.19
		Both	56.98	65.22	74.09	87.81	**88.34**

Table 2: Comparison of classifier prediction scores on all groups among the models across different tasks and attributes.

Metric	Vanilla	Sampling Baseline	Ours (I+II)	Ours (I+III)	Ours (I+II+III)
LPIPS↓	0.25 ± 0.06	0.24 ± 0.06	0.25 ± 0.06	0.24 ± 0.06	0.25 ± 0.06
MSE↓	0.06 ± 0.03	0.05 ± 0.03	0.06 ± 0.03	0.05 ± 0.03	0.06 ± 0.03

Table 3: Quantitative results for image reconstruction in the super-resolution task. Our approach does not compromise image quality.

important tool for image-to-image translation models in the presence of bias.

4.2 Generalization to a Different Architecture

Here we discuss our results when we apply our framework to pix2pix [17] on datasets other than human faces. Figure 4 shows the qualitative results among the models. For both datasets, a common pattern is that the quality of the majority does not change by much. The quality of the minority, however, varies a great deal. For 'Bags and Shoes', we can see both Vanilla and the Sampling Baseline model try to fill the gap between the body of the bag and strap. This is because most of the images from this dataset are shoes, and the contour of shoes are always filled. Therefore, to resemble the majority ground truth images, the generative model tries to fill the gap. The contrastive learning based approaches, especially (I+II+III), do not fill up the space between strap and bag as much, and show minimal changes compared to the other models. Similarly, for 'Cats and Dogs', the majority class is 'Cats,' and a frequent pattern is images having cat-like fur. As a result, both Vanilla and Sampling Baseline's outputs have cat-like fur in the minority group's ('Dog') images. However, as can be seen from this figure, our (I+II+III) model's coloring is more authentic for the 'Dogs' class. We have also quantitatively evaluated the performance of all the models by calculating FID scores between the generated images and ground truth. Table 4 shows the results. As we can see, our model achieves the lowest scores, especially for the minorities, indicating better quality of images for the selected task. For example, adding contrastive loss to re-sampling with auxiliary classifier leads to 18% reduction of overall FID for 'Bags and Shoes' dataset.

Overall, our contrastive-learning based framework leads to consistent improvement for both pSp and pix2pix (Table 2 and 4, Figure 3 and 4).

Group	Bags and Shoes					Cats and Dogs				
	Vanilla	Sampling Baseline	Ours (I+II)	Ours (I+III)	Ours (I+II+III)	Vanilla	Sampling Baseline	Ours (I+II)	Ours (I+III)	Ours (I+II+III)
Minority	202.30	135.22	140.26	130.73	**122.77**	241.98	189.35	181.29	183.04	**177.22**
Majority	233.92	159.33	151.63	152.66	**116.89**	**64.32**	70.40	76.22	65.82	68.42
Both	195.29	130.16	128.51	124.66	**105.18**	136.86	116.92	116.24	111.99	**110.7**

Table 4: FID scores show the effectiveness of our approach in a different image-to-image translation architecture, using images from different domains.

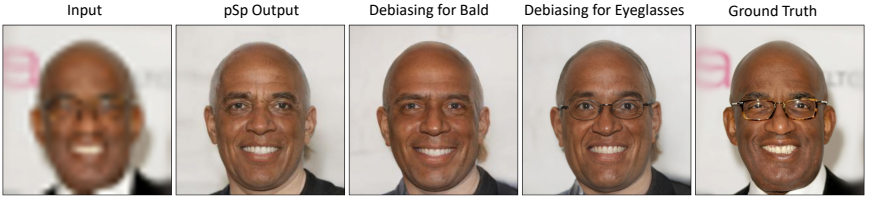


Figure 5: An example case where dual bias can appear. In this example, the ground truth image has both 'Bald' and 'Eyeglasses' attribute, and debiasing for only one attribute does not necessarily debias for the other one.

5 Discussion and Limitations

Our debiasing framework achieves better performances compared to other alternatives across different categories, different architectures, and different domains. However, so far, we have made an assumption that the bias is known, which is the most common assumption in many debiasing work [1, 9, 13, 33, 34]. This is because as long as there are attributes in the dataset, we will always be able to know if the dataset has class imbalances and whether that might lead to bias in the model. There can be various ways to measure the bias, and in Section 3.1, we explored one way of measuring it.

In this work, there is another inherent assumption that the dataset is biased towards only a single attribute/class. In reality, bias can appear in multiple attributes/classes simultaneously. For example, Figure 5 shows minority attributes, namely 'Bald' and 'Eyeglasses', appear in the same image. The figure also shows how focusing on only one attribute might not necessarily debias it for the other one (i.e. debiasing for 'Bald' does not debias for 'Eyeglasses', and vice versa). One simple, straightforward way to tackle this problem will be to merge multiple labels into single classes and apply our debiasing framework. However, doing so might not be scalable if the labels are large in number. We would like to explore this direction in future.

6 Conclusion

In this paper, we propose the new task of debiasing image-to-image translation models. Using Pixel2Style2Pixel and pix2pix, we have demonstrated that minority attributes are poorly reconstructed whenever there is an imbalance in the dataset. To solve this problem, we have proposed a novel contrastive-learning based approach to separate the latent codes of minority classes from the majority classes. From the experimental results from both pSp and pix2pix, we have shown that this contrastive learning approach, when coupled with general tricks like re-sampling and auxiliary classifiers, leads to consistent improvements across all the tasks. Our framework does not depend on any particular translation model or dataset, making our solution model and data agnostic.

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[2] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019.

[3] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of biasin face analysis algorithms. In *Deep Learning-Based Face Analytics*, pages 327–359. Springer, 2021.

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

[5] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deepfacedrawing: Deep generation of face images from sketches. *ACM Transactions on Graphics (TOG)*, 39(4):72–1, 2020.

[6] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9416–9425, 2018.

[7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[8] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[9] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

[10] Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

[11] 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*, 27:2672–2680, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[16] Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. Fairfacegan: Fairness-aware facial image-to-image translation. *arXiv preprint arXiv:2012.00282*, 2020.

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[18] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In *International Conference on Machine Learning*, pages 4721–4732. PMLR, 2021.

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[21] 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.

[22] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[24] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[26] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *arXiv preprint arXiv:1910.06809*, 2019.

[27] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[29] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[30] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2017.

[31] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan. *arXiv preprint arXiv:1805.09910*, 2018.

[32] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14083–14093, 2021.

[33] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020.

[34] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, 2018.

[35] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[36] Md Tanjim, Ritwik Sinha, Krishna Kumar Singh, Sridhar Mahadevan, David Arbour, Moumita Sinha, Garrison W Cottrell, et al. Generating and controlling diversity in image search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 411–419, 2022.

[37] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017.

[38] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[39] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[40] Yaxing Wang, Abel Gonzalez-Garcia, Luis Herranz, and Joost van de Weijer. Controlling biases and diversity in diverse image-to-image translation. *Computer Vision and Image Understanding*, 202:103082, 2021.

[41] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[42] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[43] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*, pages 506–523. Springer, 2020.

[44] Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, et al. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE transactions on medical imaging*, 39(1):188–203, 2019.

[45] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[46] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive gan: Improving data and minority coverage in generative models. In *European Conference on Computer Vision*, pages 377–393. Springer, 2020.

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[49] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.