Debiasing Image-to-Image Translation Models

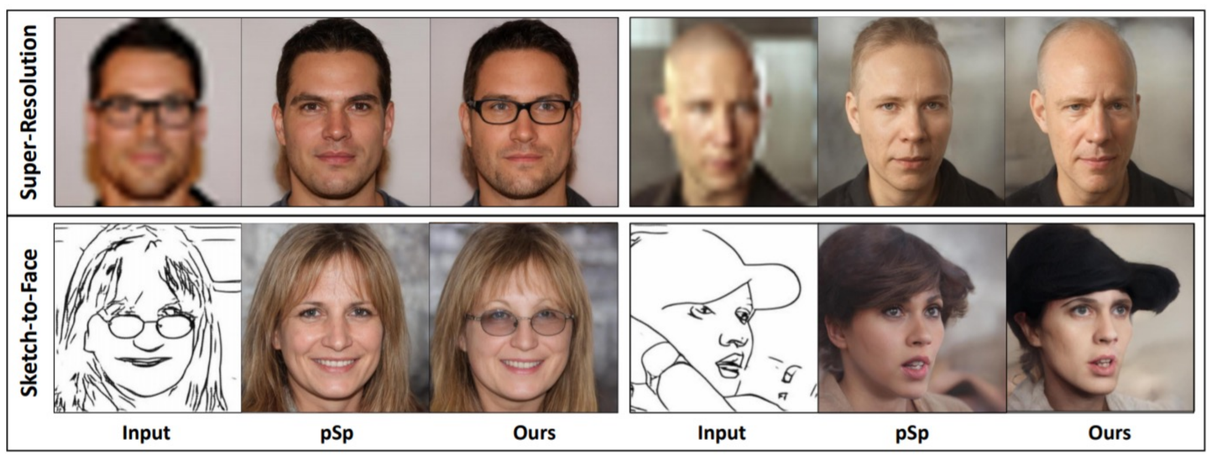
University of California San Diego

Md Mehrab Tanjim[1], Krishna Kumar Singh[2], Kushal Kafle[2], Ritwik Sinha[2], Garrison W. Cottrell[1]
[1]UC San Diego, [2]Adobe Research

Adobe Research

Overview

- In this work, we identify the bias problem in image-to-image translation (I2IT) tasks and propose a task of debiasing these models.
- We propose a novel contrastive learning-based approach which outperforms the baselines both quantitatively and qualitatively.
- We show that our model generalizes well, we apply it to multiple image-to-image translation tasks and datasets.



In this figure, we show how one of the state-of-the-art I2IT models, Pixel2Style2Pixel (pSp) [1], becomes biased against minority attributes in the CelebA-HQ dataset [2]. We also show results from our debiasing framework (Ours).

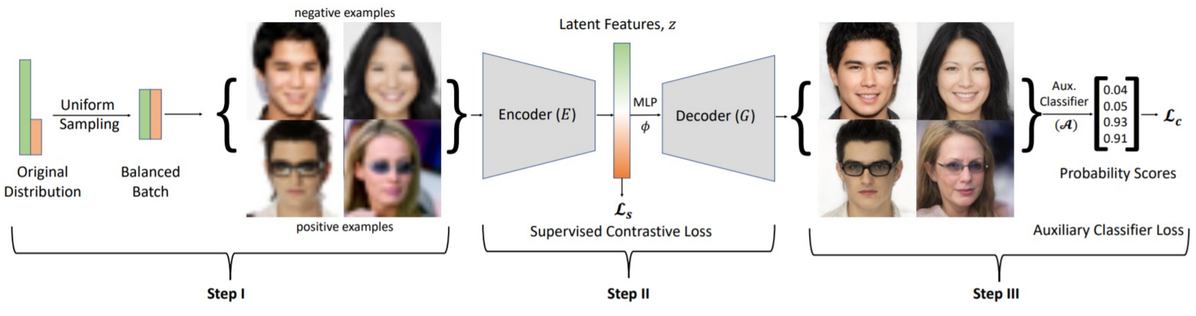
Measuring Biases

- We use CelebA-HQ [2] to measure bias in image-to-image translation tasks.
- For measuring biases quantitatively, we generate images for the super-resolution task.
- We use a ResNet152 classifier (trained on pSp training set) to calculate the F1 scores on real and generated images for measuring biases. These numbers are reported in the following Table.

Attribute	Bald	Wearing Hat	Eyeglasses	Blond Hair	Bangs	Black Hair	Male
Percentage ↓	**2.37**	**3.57**	**4.89**	17.09	18.08	21.97	36.86
F1 Score on Real	0.8142	0.8908	0.9825	0.8483	0.8756	0.8186	0.9791
F1 Score on Generated ↓	**0.7216**	**0.2500**	**0.0984**	0.8288	0.8393	0.7725	0.9653

- Lower F1 score on generated indicates bias. For debiasing, we select the first 3 attributes from Table 1, i.e.,'Bald', 'Wearing Hat', and 'Eyeglasses.'

Model



Step I: A simple way to solve class-imbalance during training is to over-sample the minority. This is our first step towards debiasing (Step I in the above Figure).

Step II: To further separate the latent codes from different groups or classes, we apply supervised contrastive loss [3]. 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,

$$\mathcal{L}_s = -\sum_{i\in\mathcal{I}} \frac{1}{|P(i)|} \sum_{p\in P(i)} \log\left(\frac{\exp(z_i . z_p)/\tau}{\sum_{a\in A(i)} \exp(z_i . z_a)/\tau}\right),$$

Here, $i \in \mathcal{I} \equiv \{1 \dots N\}$ (N is the batch size) is the index of an arbitrary sampled image. $A(i) \equiv \mathcal{I}/\{i\}$, $P(i) \equiv \{p \in \mathrm{A}(i): y_p = y_i\}$ (y is the label or class of that image), z is the hidden representation of the image, τ is the scalar temperature parameter.

Step III: To enforce the constraint on attributes further during the generation process, we use an auxiliary classifier $\mathcal{A}$ and and apply binary cross entropy loss on the generated images $G(\varphi(z_i))$:

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

where $\hat{y}_i = \mathcal{A}(G(\varphi(z_i)))$. 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

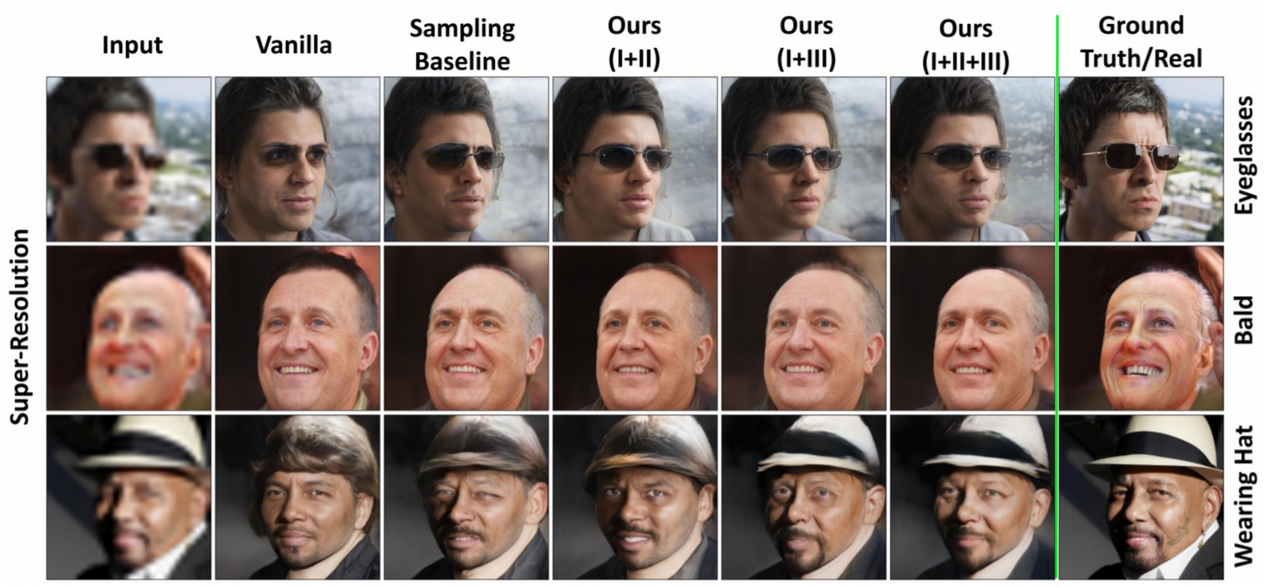
$$\mathcal{L} = \mathcal{L}_o + \lambda_s * \mathcal{L}_s + \lambda_c * \mathcal{L}_c,$$

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 and λ is the hyperparameter.

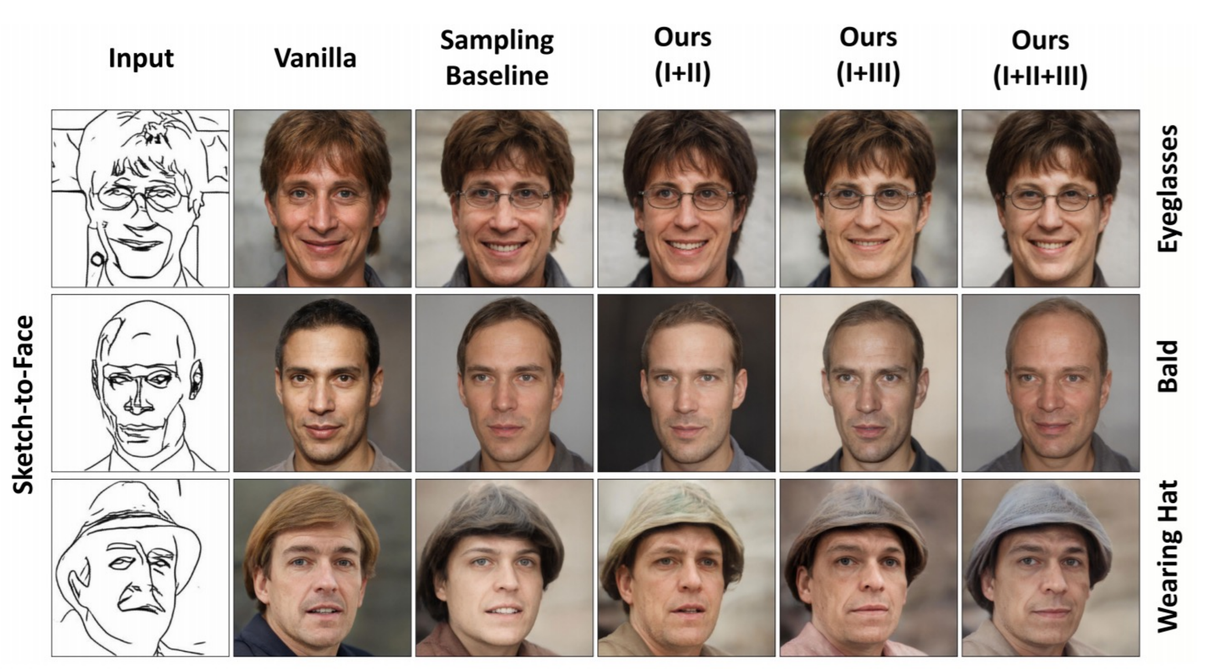
Experiments

Dataset. For experiments with human faces, we select CelebA-HQ. To test generalization to other I2IT models such as pix2pix [4], we create two synthetic datasets, namely, 'Bags and Shoes' and 'Cats and Dogs' where we select the bias ratio to 99:1 (where Bags and Dogs are a minority respectively).

Models. As there is no existing baselines, we create several baselines by applying different combinations of our debiasing steps. We also contrast our results with the original models which we refer to as Vanilla.

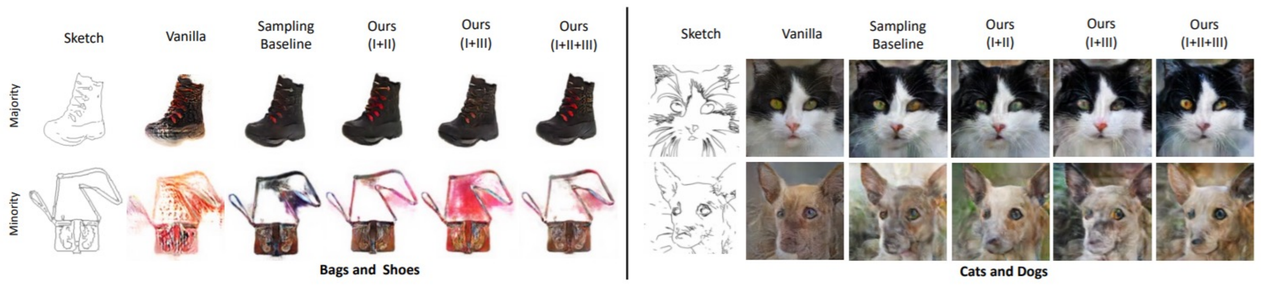


(a) Super-Resolution



(b) Sketch-to-Face

The above two figures show results of our debiasing framework. Here, we show one example for each of the considered tasks across all attributes. Our generated results better capture the attributes compared to baselines.



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

Conclusion

• In this work, we have proposed a novel contrastive-learning based approach to separate the latent codes of minority classes from the majority classes for debiasing image-to-image translation models.

• 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.

• In this work, we debiased for only a single attribute/class. As a future work, we would like to extend our framework for multiple biases.

References:

[1] Richardson et al. Encoding in style: a stylegan encoder for image-to-image translation. CVPR, 2021.
[2] Karras et al. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.
[3] Khosla et al. Supervised contrastive learning, *arXiv preprint arXiv:2004.11362* , 2020.
[4] Isola et al. Image-to-image translation with conditional adversarial networks. CVPR, 2017.