FairFaceGAN: Fairness-aware Facial Image-to-Image Translation: Supplementary Material

Sunhee Hwang
sunny16@yonsei.ac.kr
Sungho Park
qkrtjdgh18@yonsei.ac.kr
Dohyung Kim
dohkim02@yonsei.ac.kr
Mirae Do
wwwdo109@yonsei.ac.kr
Hyeran Byun
hrbyun@yonsei.ac.kr

Department of Computer Science
Yonsei University
Seoul, Republic of Korea

1 Fairness Definition

Fairness in AI can be defined as the ability to make fair decisions regarding protected attributes such as gender. In this section, we briefly describe the two of most widely used fairness metrics, which are considered in our experiment (Fair classification).

Equality of Opportunity [2] and *Equalized Odds* [6] measure whether people who need to qualify for an opportunity are likely to have the same opportunity regardless of demographic group. Formally, *Equality of Opportunity* is defined so that different gender groups have the same true positive rates for the target attribute as follows:

$$\mathcal{P}(\hat{Y}=1|p=0,Y=1)=\mathcal{P}(\hat{Y}=1|p=1,Y=1), \tag{1}$$

where p, Y, $\hat{Y} \in \{0,1\}$ denote the protected attribute (gender), the target attribute, and the prediction respectively.

Unlike *Equality of Opportunity*, which only considers true positive rate parity, *Equalized Odds* is defined by considering the true and false positive rates of different gender groups as follows:

$$\begin{aligned}&\mathcal{P}(\hat{Y}=1|p=0,Y=1)=\mathcal{P}(\hat{Y}=1|p=1,Y=1)\ \textit{and}\\&\mathcal{P}(\hat{Y}=1|p=0,Y=0)=\mathcal{P}(\hat{Y}=1|p=1,Y=0).\end{aligned} \tag{2}$$

2 Protected attributes related bias on CelebA dataset

We present how manually selected target attributes are biased in terms of demographic groups. Table 2 shows the Pearson correlation between protected attributes (*i.e.*, Male,

Table 1: Pearson correlation between manually selected five attributes and protected attributes (Male, Young) on CelebA dataset [4].

Attribute	Male	Young
Blond Hair	-0.31	0.06
Bald	0.3	-0.24
Bags Under Eyes	0.18	-0.20
Big Nose	0.37	-0.29
Attractive	-0.4	0.39

Young) and target attributes (*i.e.*, Blond Hair, Bald, Bags Under eyes, Big Nose, Attractive) on CelebA dataset [4]. Blond Hair and Attractive have a negative correlation for male and positive correlation for Young, and the others have the opposite correlation. On the other hand, we do not consider a race-related label since there is no label on CelebA dataset.

3 FairFaceGAN

Network FairFaceGAN consists of an encoder-decoder generator, a discriminator, and two protected attribute classifiers (PACs). The generator consists of two down-sampling convolutional layers followed by five residual blocks, and two deconvolutional layers. We use PatchGAN discriminator [3], which consists of six down-sampling convolutional layers. Where, we take outputs of the encoder to PACs, after flatten convolutional features. Two PACs are consists of three fully connected layers.

Additional results We illustrate additional image translation results compared with StarGAN [1] and FixedPointGAN [5] in Figure 1, 2, 3, 4, 5, and 6. $+$ denotes without a target attribute into with the target attribute, where $-$ indicates the opposite case.

Parameters We present the details of parameters to train our FairFaceGAN as shown in Table 2.

Table 2: Parameter Details.

Parameters	Value
Batch Size	16
Reconstruction Loss (Different Domain)	10
Reconstruction Loss (Same Domain)	11
Auxiliary Classifier Loss	1
Fair Representation Loss (FRL)	0.001
Protected Attribute Distance Loss (PADL)	2
Perceptual Loss (Style)	0.025
Perceptual Loss (Content)	0.01

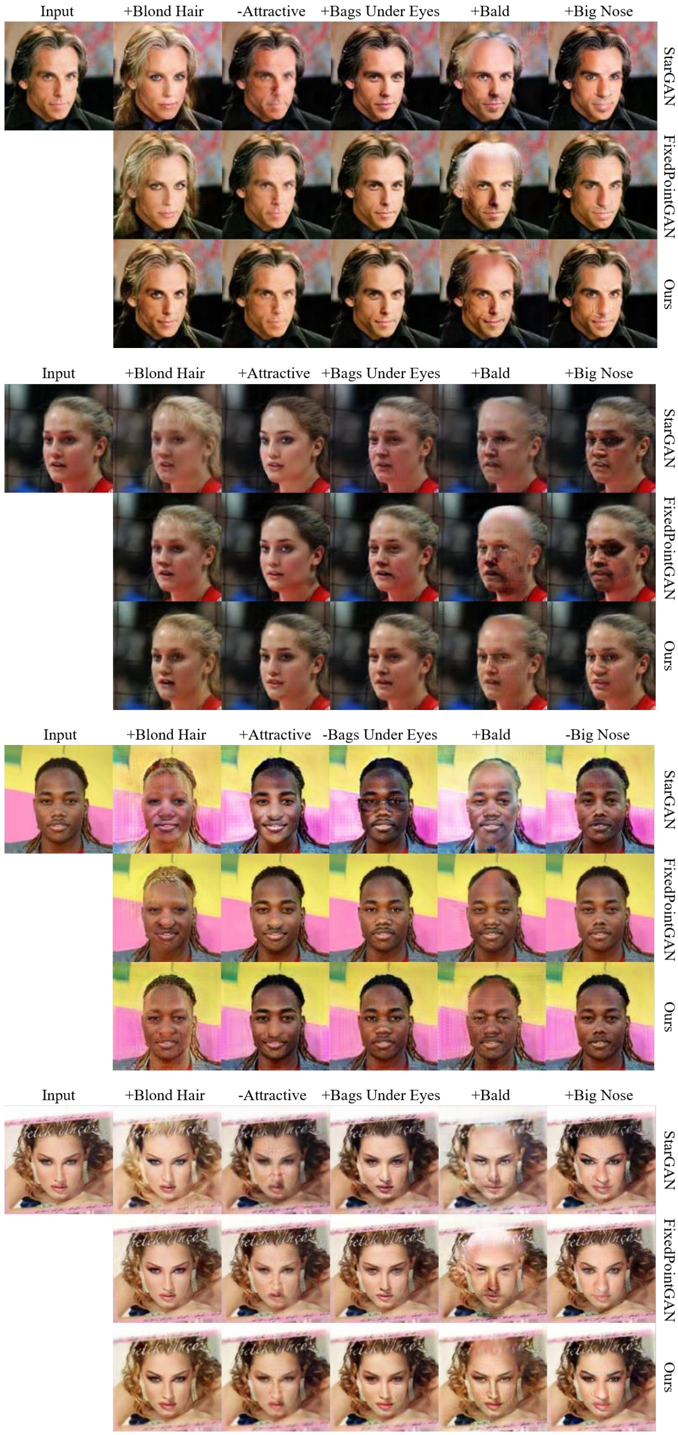

Figure 1: Image-to-Image translation results compared with StarGAN [1] and FixedPoint-GAN [5].

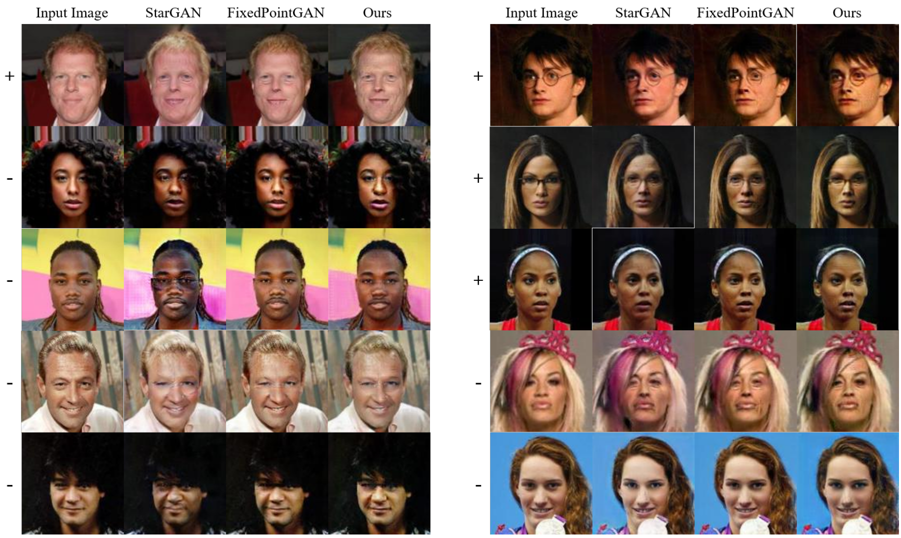

Figure 2: Results of inversion for the attribute *Bags Under Eyes*.

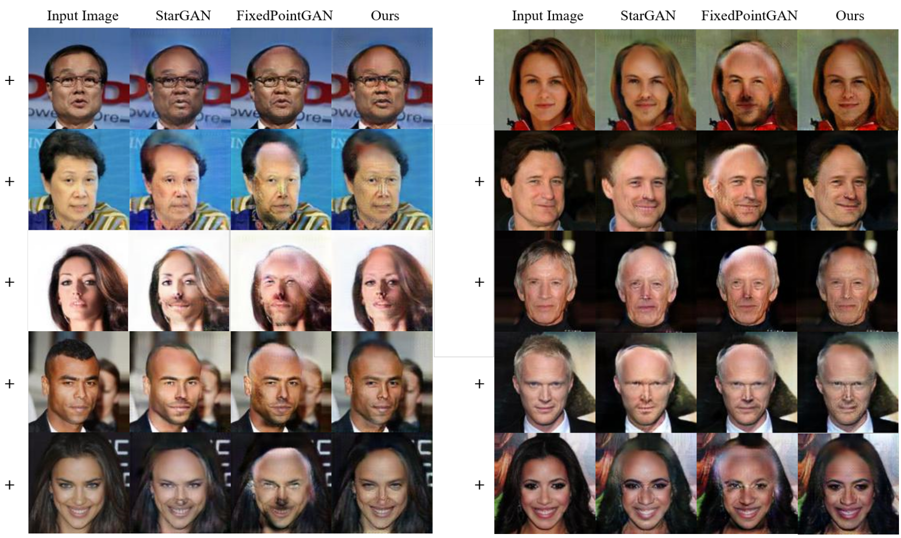

Figure 3: Results of inversion for the attribute *Bald*.

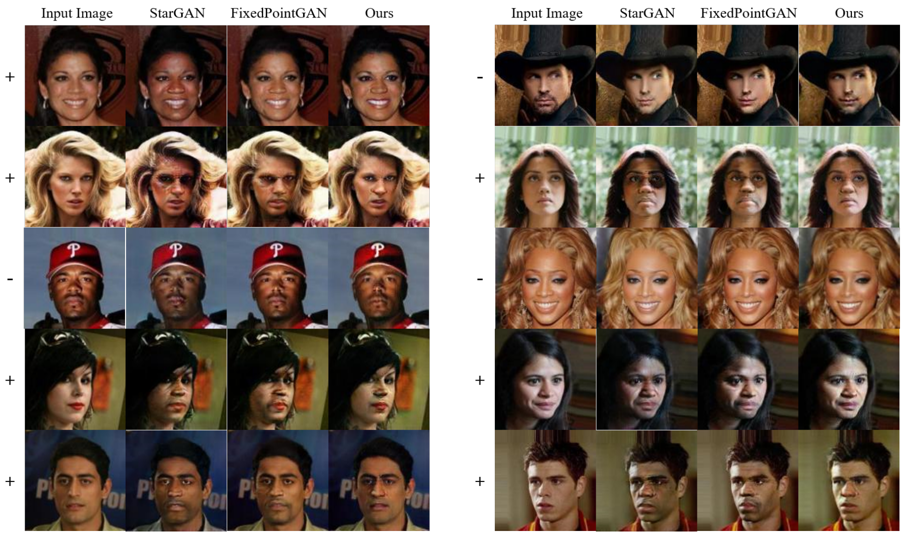

Figure 4: Results of inversion for the attribute *Big Nose*.

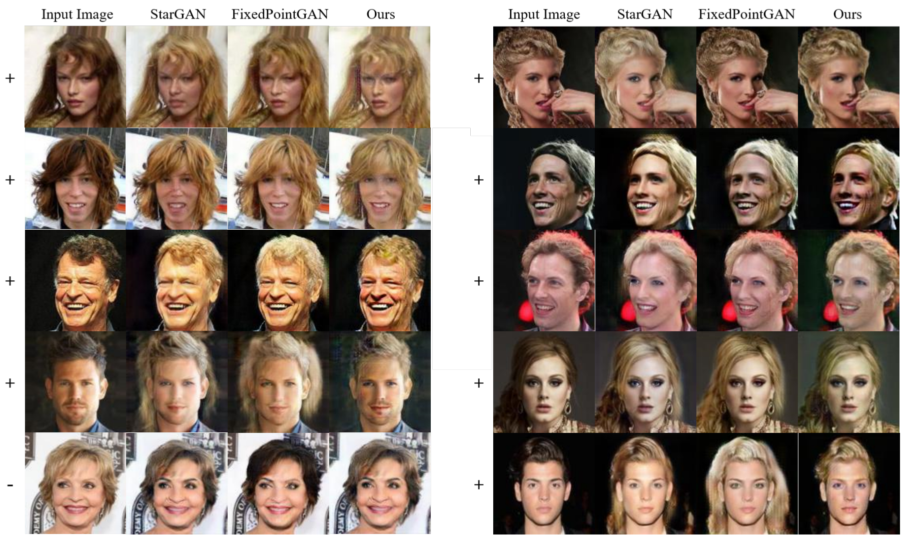

Figure 5: Results of inversion for the attribute *Blond*.

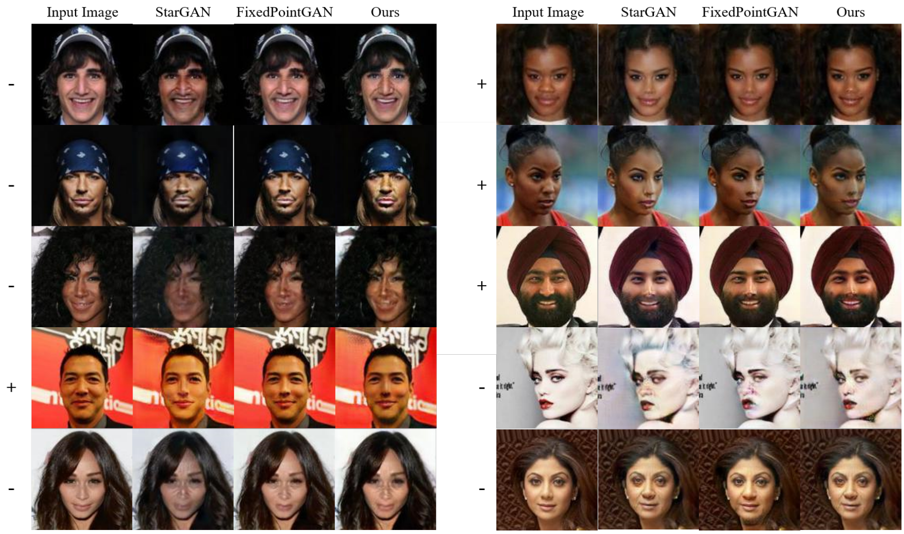

Figure 6: Results of inversion for the attribute *Attractive*.

References

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[2] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[3] 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.

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[5] Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 191–200, 2019.

[6] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.