Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-Attribute Transfer
Supplementary Material

1 Additional qualitative results

In this first section, we present additional results as a means of further validating our claims in the main paper.

Attribute Transfers We demonstrate qualitatively the generalizability of the method Fig. 3 by translating the images from the KANFace test set using the model trained on CACD. Despite being trained on a completely different dataset, we can still translate these images from KANFace to realistic examples of the target class.

Non-face datasets We show here two experiments to showcase the applicability of our method in different domains. Firstly, we introduce two additional modes of variation into the Morpho-MNIST dataset [1]: 'color', and 'thickness'. We then train our model to transfer these two attributes simultaneously, showcasing the results of the transfers in Fig. 2, where thickness and color are transferred in a realistic manner.

In Fig. 1 we also show an experiment on the CompCars dataset [3]. We threshold each images' average pixel intensity to introduce a 'daytime' attribute, and then train a model to transfer both color (red, white and black) and the time of day (day and night) simultaneously. Our model learns a continuous representation of car color, leading to the ability to translate various intensities of each attribute, despite only being shown discrete labels at train-time. Indicatively, our models generates different shades of the annotated colors (e.g., magenta) as well as continuous illumination.

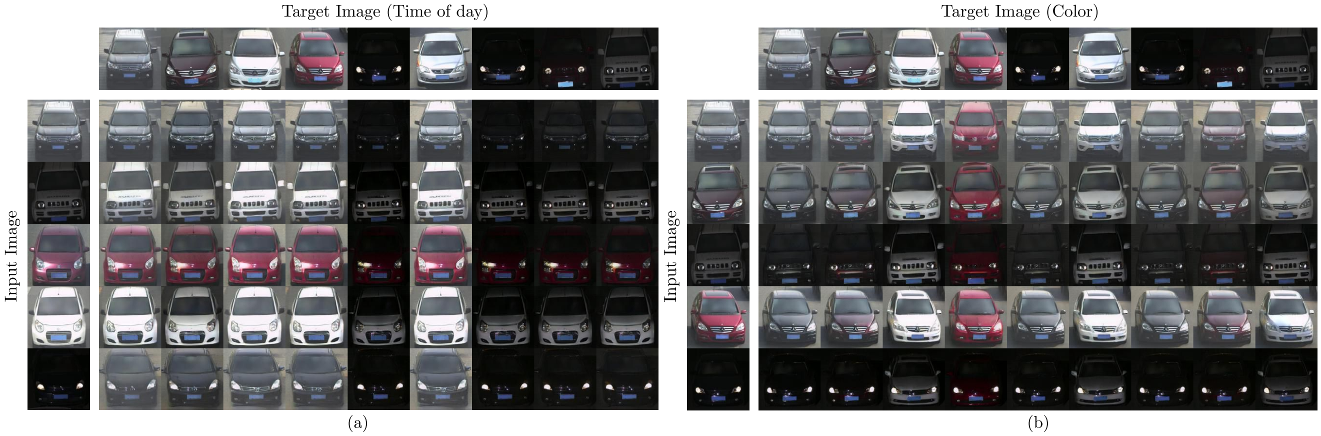
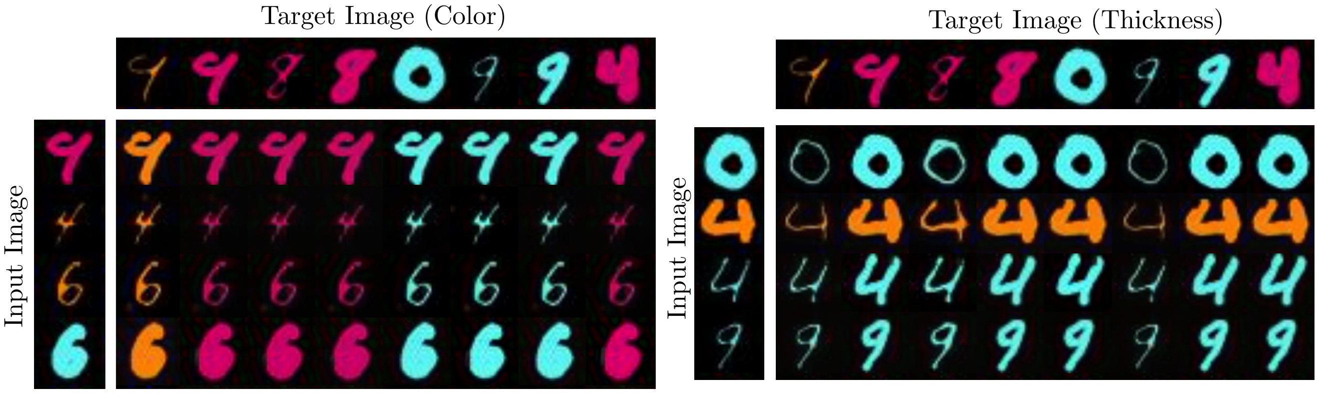


Figure 1: Additional results on multiple attribute transfer–for (a) time of day, and (b) car color–on the CompCars [3] dataset using our method.

2 Additional quantitative results

Here we show the full quantitative results for mitigating skin-tone bias for the gender recognition task. As in the main paper, we augment the MORPH dataset by translating every training image to every other skin-tone and age combination. We then train a VGG16 classifier on these synthetic sets, and report the *Equality of Opportunity*



(a) Color Transfer

(b) Thickness Transfer

Figure 2: Additional results on multiple attribute transfer on the Morpho-MNIST [1] dataset using our method.

for every age and skin-tone combination in Fig. 4. The (s1, a4) row corresponds to the 60+ dark-skinned female category, where the bias is more prevalent. Training on augmentations from our method results in notably better EO than that from the baselines. For reproducability purposes, we also provide the precise support of each dataset used in Table 1.

	a0	a1	a2	a3	a4	g0	g1	s0	s1
CACD	2136	29671	41840	31246	987	53281	53719	-	-
MORPH	1537	16198	17841	6725	249	6762	37345	8450	34063
UTKFace	3391	5956	3977	2416	1788	9315	8398	7491	3244

Table 1: The support for each attributes for all datasets (we note that there are no skin tone labels present in the CACD dataset).

3 Ablation studies

For the reconstruction and adversarial losses, we choose the default values used in previous works [2]. We also conduct an ablation study to showcase the sensitivity of our model to hyperparameter choice. In particular, we train 9 different combinations of these two hyperparameters and present qualitative results in Fig. 5a. For the rank of the CP decomposition in the multiplicative fusion layers, we choose the value via a grid search. We show in Fig. 5b the results when we vary the rank of the decomposition, and the resulting transfers.

4 Detailed Network Architecture

The specifics of our networks, including our choice of layers and activation functions, can be found in Table 6 and Table 7.

References

[1] Daniel Coelho de Castro et al. “Morpho-MNIST: Quantitative Assessment and Diagnostics for Representation Learning”. In: *J. Mach. Learn. Res.* 20 (2019), 178:1–178:29.

[2] Ming-Yu Liu et al. “Few-Shot Unsupervised Image-to-Image Translation”. en. In: (May 2019). URL: https://arxiv.org/abs/1905.01723v2 (visited on 03/11/2020).

[3] Linjie Yang et al. *A Large-Scale Car Dataset for Fine-Grained Categorization and Verification.* 2015. arXiv: 1506.08959 [cs.CV].

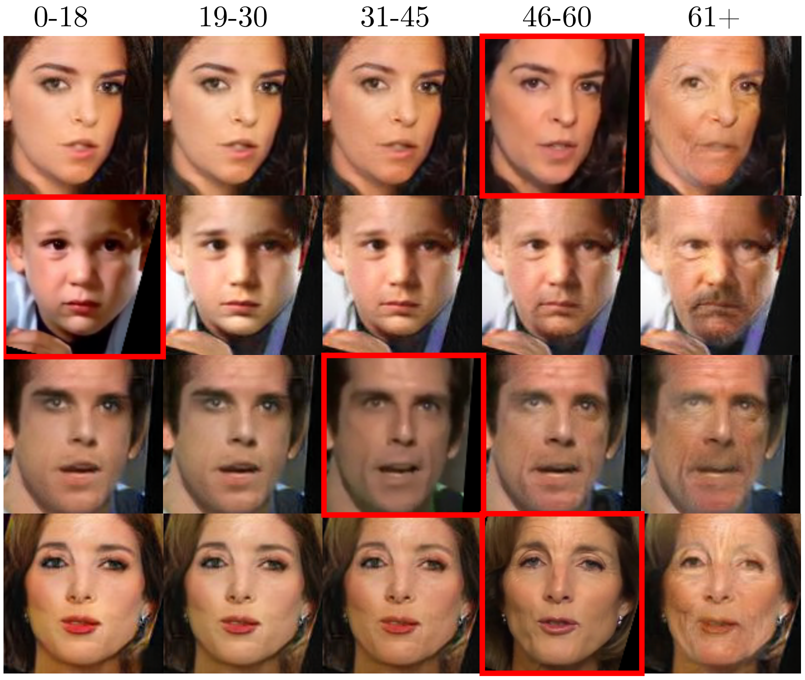


Figure 3: Age translations on the KANFace test set, using the model trained on CACD. The red squares picture the input images in the location of their class labels.

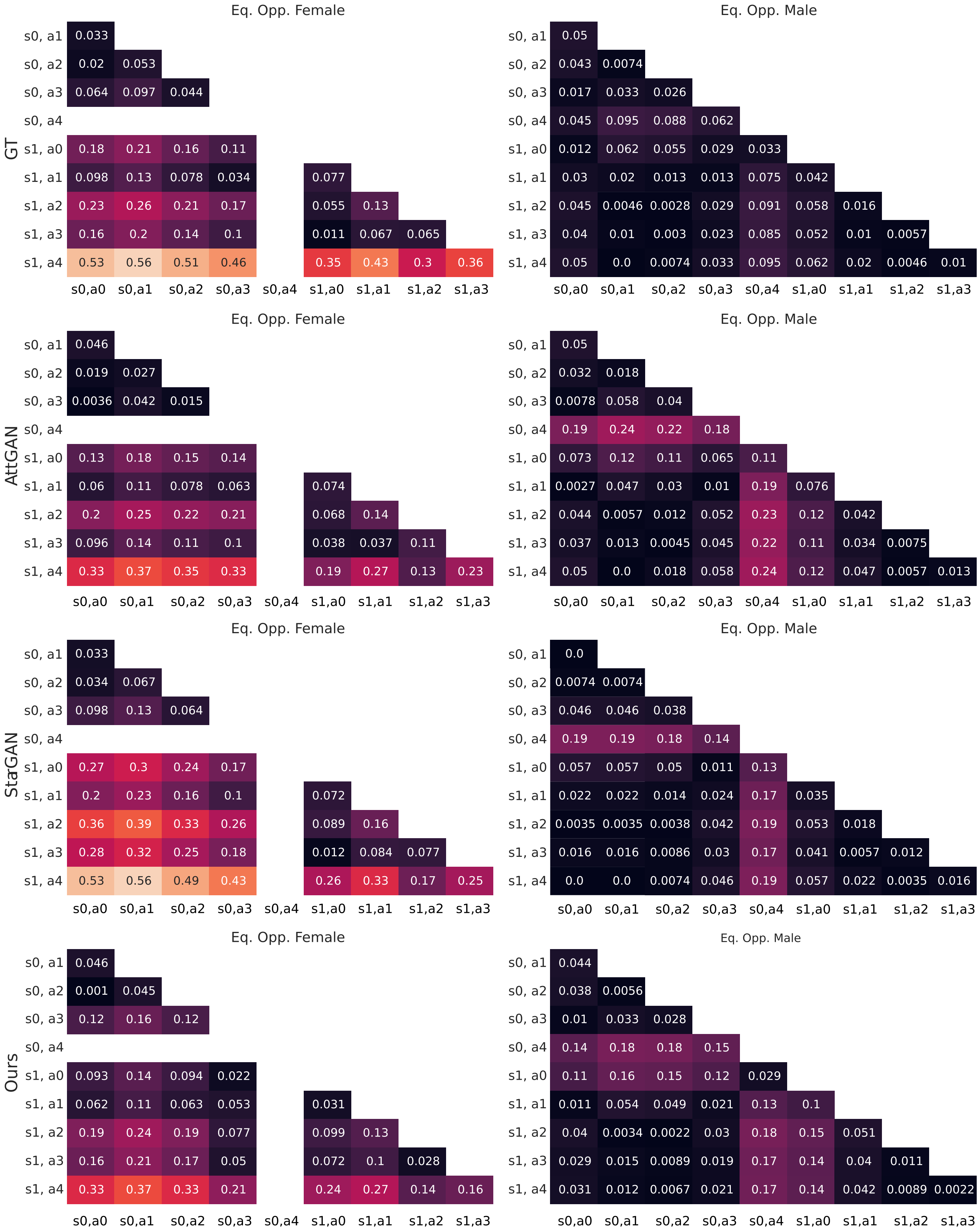
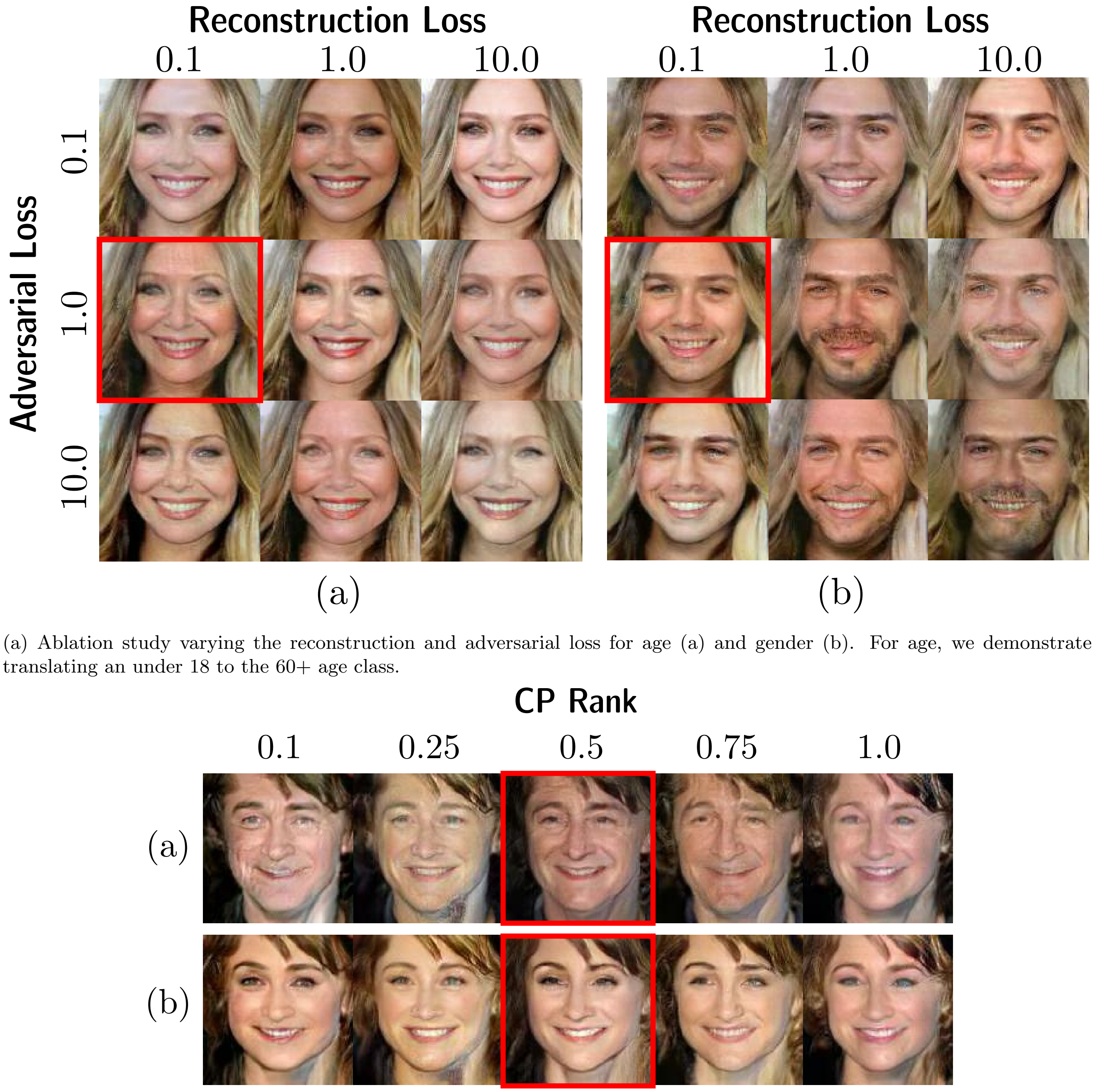


Figure 4: Equality of opportunity for each skin-tone and age combination for the MORPH test set, when trained on the baselines' synthetic images. 'si, aj' denotes skin tone class i and age class j.



(a) Ablation study varying the reconstruction and adversarial loss for age (a) and gender (b). For age, we demonstrate translating an under 18 to the 60+ age class.

(b) Ablation study varying the CP rank of the decomposition for the tensor modeling the higher-order interactions of multiple attributes' latent style vectors for age (a) and gender (b).

Figure 5: Ablation studies varying various hyperparamters of our model. The red box indicates the chosen value for the final models.

Section	Name	Dimensions (In)	Dimensions (Out)	Layers
Encoder	e1	$(b, 128, 128, 3)$	$(b, 128, 128, 64)$	Conv2d(f=64, k=7, s=1) → IN → ReLu
	e2	$(b, 128, 128, 64)$	$(b, 64, 64, 128)$	Conv2d(f=128, k=4 s=2) → IN → ReLu
	e3	$(b, 64, 64, 128)$	$(b, 32, 32, 256)$	Conv2d(f=256, k=4 s=2) → IN → ReLu
Residual Blocks (×6)	z	$(b, 32, 32, 256)$	$(b, 32, 32, 256)$	m-AdaIN → Conv2d(f=256, k=3 s=1) → ReLu → m-AdaIN → Conv2d(f=256, k=3 s=1) → ReLu
Decoder	d5	$(b, 32, 32, 256)$	$(b, 64, 64, 128)$	ConvT2d(f=128, k=4 s=2) → m-AdaIN → ReLu
	d6	$(b, 64, 64, 128)$	$(b, 128, 128, 64)$	ConvT2d(f=64, k=4 s=2) → m-AdaIN → ReLu
	d7	$(b, 128, 128, 64)$	$(b, 128, 128, 3)$	ConvT2d(f=3, k=7 s=1)

Figure 6: Encoder and Decoder architecture: b denotes the mini-batch size, and 'm-AdaIN' refers to the proposed multilinear extension of the Adaptive Instance Normalization operation.

Section	Name	Dimensions (In)	Dimensions (Out)	Layers
Discriminator	d1	$(b, 128, 128, 3)$	$(b, 128, 128, 64)$	Conv2d(f=64, k=7, s=1) → LeakyReLu
	d2	$(b, 128, 128, 64)$	$(b, 64, 64, 128)$	Conv2d(f=128, k=4 s=2) → LeakyReLu
	d3	$(b, 64, 64, 128)$	$(b, 32, 32, 256)$	Conv2d(f=256, k=4 s=2) → LeakyReLu
	logits	$(b, 32, 32, 256)$	(b,a)	Flatten → FC(a)

Figure 7: Discriminator architecture: b again denotes the mini-batch size, and a denotes the number of classes for the attribute the discriminator governs.