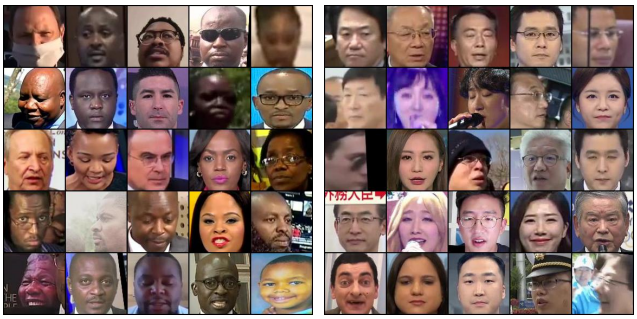
Supplementary Material
How to Boost Face Recognition with StyleGAN?

Artem Sevastopolsky[1] Yury Malkov[2,*] Nikita Durasov[3]
Luisa Verdoliva[1,4] Matthias Nießner[1]

[1] Technical University of Munich, Germany [2] Twitter, US
[3] École polytechnique fédérale de Lausanne, Switzerland
[4] University Federico II of Naples, Italy

A. Data collection

To scrape AfricanFaceSet-5M and AsianFaceSet-3M, we followed the same pipeline. First, a list of channels with predominantly African *(List 1)* and Asian *(List 2)* demographics is provided (see below). List of IDs of all videos released in each of the channels is collected via scrapetube library. Typical number of videos in a channel is 50K–150K. Afterwards, the videos are downloaded one-by-one in the highest available quality with a limit of 20 min per video via pytube in 36 parallel threads. The time limit is specified in the video URL, which allows not to download excessive parts of videos. One frame per every $P = 5$ is extracted via ffmpeg, and faces are extracted using MTCNN [9] (mtcnn-pytorch implementation) and are aligned by landmarks via `cv2.warpAffine` from OpenCV. To ensure the correctness of the MTCNN face detector, we impose constraints in terms of the minimal face size in the original image (100×100 px) and value of 0.9 set for all the detector thresholds. Example pictures are provided in Fig. 4a and Fig. 4b.



(a) Examples from AfricanFaceSet-5M (b) Examples from AsianFaceSet-3M

Figure 1: Random samples from the collected AfricanFaceSet-5M and AsianFaceSet-3M datasets.

YouTube channels used for the collection of AfricanFaceSet-5M *(List 1)*:

- Africa24 https://www.youtube.com/user/Africa24
- African Glitz https://www.youtube.com/c/AfricanGlitzTV
- Arise News https://www.youtube.com/c/AriseNewsChannel
- BBC News Africa https://www.youtube.com/c/BBCAfrica
- Best African TV https://www.youtube.com/channel/UCz2lohZJpkfOkvXVyu_6wng
- CGTN Africa https://www.youtube.com/c/cgtnafrica
- Channels Television https://www.youtube.com/c/ChannelsTelevision
- DStv https://www.youtube.com/dstv
- eNCA https://www.youtube.com/c/encanews
- Eye Witness News https://www.youtube.com/c/EyeWitnessNewsBahamas
- Guardian Nigeria https://www.youtube.com/c/GuardianNigeriaOfficial
- Legit TV https://www.youtube.com/c/LegitTV
- News Central TV https://www.youtube.com/c/NewsCentralTVafrica
- Newzroom Africa https://www.youtube.com/channel/UCQMML3hAsx-Mz9j9ZN0tThQ
- One Africa TV https://www.youtube.com/c/OneAfricaTelevision
- Plus TV Africa https://www.youtube.com/c/PlusTVAfrica
- Roots TV https://www.youtube.com/c/RootsTVCommunity
- SABC News https://www.youtube.com/sabcnews
- TVC News Nigeria https://www.youtube.com/c/tvcnewsnigeria
- Voice TV Nigeria https://www.youtube.com/c/VoicetvNigeria

- The Walk https://www.youtube.com/c/TheWalkk (without the 20 min video limit, since the channel contains long walking tours over cities)
- Kenya Citizen TV https://www.youtube.com/c/kenyacitizentv
- Africa News https://www.youtube.com/c/africanews
- African Tigress https://www.youtube.com/c/AFRICANTIGRESS

YouTube channels used for the collection of AsianFaceSet-3M *(List 2)*:

- Asian Boss https://www.youtube.com/c/AsianBoss
- CCTV Video News Agency https://www.youtube.com/c/CCTVVideoNewsAgency
- China Daily 中国日报 https://www.youtube.com/channel/UCahujLjSL34EPNxtwKRi_vg
- China Live 直播中国 https://www.youtube.com/c/chinanews
- China Matters https://www.youtube.com/c/ChinaMatters
- CNA https://www.youtube.com/user/channelnewsasia
- Discovery Channel Southeast Asia https://www.youtube.com/c/DiscoveryChannelSEAsia
- New China TV https://www.youtube.com/c/ChinaViewTV
- Nikkei Asia https://www.youtube.com/user/NikkeiAsia
- South China Morning Post https://www.youtube.com/c/SouthChinaMorningPost
- Tencent Video https://www.youtube.com/channel/UCQatgKoA7lylp_UzvsLCgcw
- Top Korean News https://www.youtube.com/c/TopKoreanNews
- ANNnewsCH https://www.youtube.com/user/ANNnewsCH
- Arirang News https://www.youtube.com/c/ArirangCoKrArirangNEWS
- Ask Japanese https://www.youtube.com/c/AskJapanese
- CCTV Video News Agency https://www.youtube.com/c/CCTVVideoNewsAgency
- CCTV中国中央电视台 https://www.youtube.com/c/cctv
- CCTV今日说法官方频道 https://www.youtube.com/user/jinrishuofa
- CCTV挑战不可能官方频道 https://www.youtube.com/channel/UC3HLhJGcc_0Vse2UncGnxcQ
- CCTV春晚 https://www.youtube.com/c/cctvgala
- CCTV电视剧 https://www.youtube.com/channel/UC7Vl0YiY0rDlovqcCFN4yTA
- CCTV社会与法 https://www.youtube.com/c/Internationalcntv
- CCTV科教 https://www.youtube.com/user/kejiaotv
- CCTV纪录 https://www.youtube.com/user/documentarycntv
- DKDKTV https://www.youtube.com/c/DKDKTV
- Hi China https://www.youtube.com/c/CCTVcomInternational
- KBS WORLD TV https://www.youtube.com/c/kbsworldtv
- KOREA NOW https://www.youtube.com/c/KOREANOWyna
- Live Japan https://www.youtube.com/channel/UCW879NMJHIvKspfOg3H8OsQ
- NHK WORLD-JAPAN https://www.youtube.com/c/NHKWORLDJAPAN
- Nippon TV News 24 Japan https://www.youtube.com/c/NipponTVNews24Japan
- ShanghaiEye 魔都眼 https://www.youtube.com/c/Kankanewsbilingual
- The Japan Times https://www.youtube.com/user/thejapantimes
- The Thaiger https://www.youtube.com/c/TheThaiger
- Tokyo Explorer https://www.youtube.com/c/TokyoExplorer
- VisitSeoul TV https://www.youtube.com/c/VisitSeoulTV
- Walk East https://www.youtube.com/c/WalkEast
- 新 TVB NEWS official https://www.youtube.com/channel/UC_ifDTtFAcsj-wJ5JfM27CQ

The examples of positive and negative pairs of both publicly available RFW and newly assembled RB-WebFace are shown in Fig. 2. As shown, both datasets feature challenging pairs, however in RB-WebFace the evaluation protocol is different: for RB-WebFace, we calculate TPR given predefined FPR for a small number of positive pairs and a large number of negative pairs, while for RFW, simple calculation of accuracy is possible, since the number of positive pairs and negative pairs is the same. Using all possible negatives for RB-Webface allows to reduce the potential selection bias, as there is no longer any need to select challenging negatives by a face recognition network.

B. Additional comparisons

TPR and FPR values at all thresholds. The comparison in Fig. 3 is a graphical representation of quality of the methods on RB-WebFace described in Table 3 in the main paper text. Here we showcase the same data not for the fixed FPR but for all FPR in the form of plots, obtained by sweeping a threshold.

Effect of the collected prior datasets vs. using FFHQ. Here we evaluate whether it makes sense to employ larger and more diverse unlabeled data collections in the pretraining stage by comparing $\mathcal{D}^{prior}$ = AfricanFaceSet ∪ AsianFaceSet to $\mathcal{D}^{prior}$ = FFHQ. Relative to baseline, pretraining on FFHQ helps all ethnicities and mostly Caucasian, which is indeed the predominant group in FFHQ.

Dataset name	# people	# pos pairs	# neg pairs	# subgroups	neg pairs not by facerec	preferred protocol
IJB-A	500	$\leq$ 23K	100K	1	✓	ROC curve
DemogPairs	600	91K	19M	6	✓	ROC curve
BFW	800	240K	681K	8	✓	ROC curve
RFW	12K	12K	12K	4	✗	LFW-like
RB-WebFace	**72K**	**360K**	**648M**	4	✓	ROC curve

Table 1: Overview of the existing publicly available datasets of pairs used to evaluate face recognition accuracy. Since the appearance of LFW [4], many test sets consisting of the same number positive (same person) pairs and negative (similar people) pairs have been proposed. The RFW dataset [7] is compiled from MS-Celeb-1M [2] in a similar fashion for the purpose of fairness estimation of a trained face recognizer and considered the standard benchmark for fairness. We propose a new testing set for the fairness estimation – *RB-WebFace* – comprising a partition of recently released WebFace-42M, which addresses two issues of RFW. First, we use all negative pairs instead of their subset selected by a pretrained face recognition network that can be potentially introduce selection bias. Second, the dataset contains much larger number of pairs. As we show, RB-WebFace is also a harder (less saturated) benchmark.

	RFW, accuracy %, ↑					
	Cauc.	African	Asian	Indian	avg ↑	std ↓
ArcFace R-50	96.00	94.00	93.08	94.48	94.39	1.06
+ $\mathcal{D}^{prior}$ (Afr+Asian)	96.52	**95.00**	**93.90**	**94.93**	**95.09**	**0.94**
+ $\mathcal{D}^{prior}$ (FFHQ)	**96.58**	94.42	93.65	94.53	94.80	1.09

Table 2: Comparison to pretraining on FFHQ dataset.

Comparison of the StyleGAN encoder architectures. We provide the ablation over encoder training strategies in Table 3. There's no specific strategy that yields the best result across all groups, but by avg and std, pSp is the best-performing choice of architecture.

Method to train	RFW, accuracy %, ↑					
the R-50 encoder	Cauc.	African	Asian	Indian	avg ↑	std ↓
pSp [5]	96.52	**95.00**	93.90	94.93	**95.09**	**0.94**
e4e [6]	96.40	94.08	**94.10**	**95.05**	94.91	0.95
ReStyle [1]	**96.67**	94.43	93.83	94.80	94.93	1.06

Table 3: Comparison of StyleGAN encoders to use in Stage 2 and 3 (Afr+Asian). ReStyle is based on cascaded prediction and, in this experiment, iterates through pSp base architecture three times per pass.

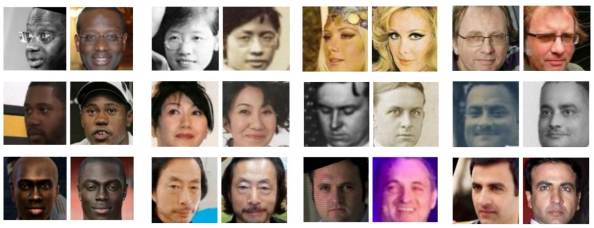
Application for gender classification. The demonstrated encoder-based pretraining technique is also applicable to other tasks. To show that, we conduct a simple experiment where the pSp R-34 encoder, pretrained in Stage 2, is fine-tuned for gender classification, not face recognition task. As a labeled dataset, we take Kaggle 200K gender recognition from CelebA dataset, and fine-tune the encoder on it with binary cross-entropy loss. Just like for our primary downstream task, the results indicate that the quality boost is especially prominent for a limited amount of labeled data. When trained on 1% of the labeled dataset, **94.42%** accuracy is achieved with our pretraining and **75.17%** without it. For 10% of the labeled dataset, **97.04%** accuracy with our pretraining vs **93.47%** without is achieved. For the full dataset, the quality difference was saturated. In this experiment, we freeze the encoder for the first 8 epochs when training on 1% of the labeled dataset (both w/ and w/o pretraining) to avoid SGD convergence issues.

Prior datasets filtering. We found that applying strict ethnicity filtering via consensus-based classifier (see Subsec. C.1) on AfricanFaceSet and AsianFaceSet removes around 30% of the collected faces. Unlike the case when no filtering is performed (Table 2 in the main paper text), pretraining on the filtered data results in more evident same-race improvement (i.e., pretraining on African helps more on RFW-African benchmark, and pretraining on Asian helps more on RFW-Asian) – see Table 4.

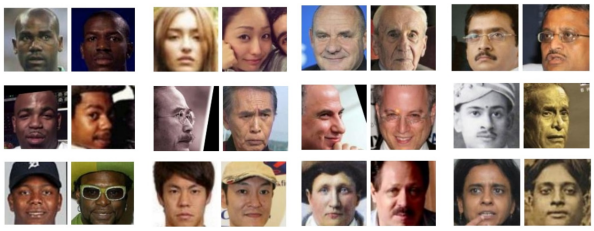
	RFW, accuracy %, ↑			
	Cauc.	African	Asian	Indian
Baseline (ArcFace R-50)	96.00	94.00	93.08	94.48
Baseline + $\mathcal{D}^{prior}$ (African-**F**)	96.10	**94.93**	93.70	95.27
Baseline + $\mathcal{D}^{prior}$ (Asian-**F**)	96.70	94.53	**94.23**	94.75

Table 4: Comparison on RFW with filtered (-**F**) prior datasets.

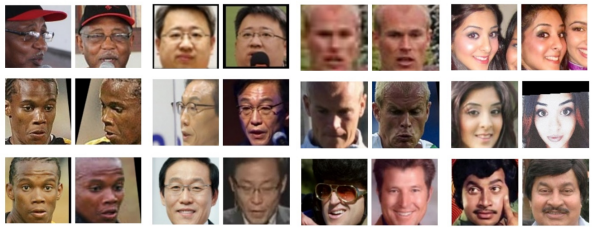
Even with the filtering applied, certain improvement for one ethnicity can be observed even after pretraining on another ethnicity (e.g., Asian-**F** also helps on African). The improvement can probably be attributed to transfer learning of general face features/conditions/geometry independent of the subject ethnicity. For instance, FFHQ pretraining, predominantly Caucasian, also aids in recognizing other ethnicities (see Table 2).



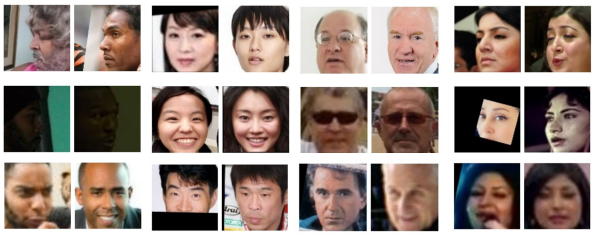
(a) RFW positive pairs examples



(b) RFW negative pairs examples



(c) RB-WebFace positive pairs examples



(d) RB-WebFace negative pairs examples

Figure 2: Examples of positive and negative pairs on RFW and newly assembled RB-WebFace (partition on WebFace-42M).

C. Implementation details

C.1. RB-WebFace

Here we provide additional information about the construction of the RB-WebFace protocol, which is done in several stages. First, images from WebFace-42M are processed by an ethnic group classifier pretrained on BUPT-BalancedFace that makes the initial judgment of whether the person belongs to the African, East Asian, Indian, or Caucasian group. Since WebFace-42M contains several images per each of its 2M people, we apply the consensus algorithm to make the classifier's decision more confident about the person's ethnic group. Specifically, we consider the person belonging to the ethnic group E, $E \in \{1, .., 4\}$, if there are at least 14 photos of this person in the dataset, and the ethnic group classifier predicts the group E for at least 80% of their photos (not more than 20 random photos of the person are considered). Subsequently, N people from each group are selected and M positive pairs are constructed from them. A set of negative pairs is constructed as a compilation of all possible distinct pairs of N pictures (1 random image of each person). We used the maximal possible value for $N = 18000$ (the lowest number of people across 4 ethnic groups, for which the race classifier was sure about the race). For each person, five positive pairs are selected, resulting in 90 K positive and $\sim$162 M negative pairs per ethnic group. Since a pretrained face recognition network can potentially induce bias in the selection of negative pairs, we deliberately make use of all the possible $\mathcal{O}(N^2)$ negative pairs.

C.2. Training procedure

This section reveals a number of implementation details not covered in the main paper text.

For the first pretraining stage (StyleGAN2-ADA training), we used the stylegan2-ada-lightning implementation and trained it with the following hyperparameters:

latent dim	# layers (z $\rightarrow$ w)	G lr	D lr
512	8	0.002	0.00235

λ_{gp}	λ_{plp}	ada_start_p	ada_target
4.0	2.0	0.0	0.6

The number of samples seen during training is set to 8 million, which roughly corresponds to the observed number of iterations when FID reconstruction score stops decreasing during fitting. The choice of $\lambda_{gp} = 4$ and an 8-layer mapping network is relatively unconventional and proved best in our setting. The resolution of output images was set to 128×128. Training was performed on 4 NVIDIA RTX 2080 Ti GPUs with 11 GB memory size each, with mini-batch size of 32 and without mixed precision.

For the second pretraining stage (pSp encoder training), we used the restyle-encoder implementation, manually adapted for the use with StyleGAN2-ADA generator. ReStyle can be seen as a generalization of pSp with only one cascade step. The hyperparameters:

L_2 weight λ_1	LPIPS weight λ_2	ID weight λ_3	reg weight λ_4
1	0.8	0	0

ID weight λ_3 was disabled on purpose to avoid the identity information leaking into the encoder during training, so that the experiment is fair. The input images are

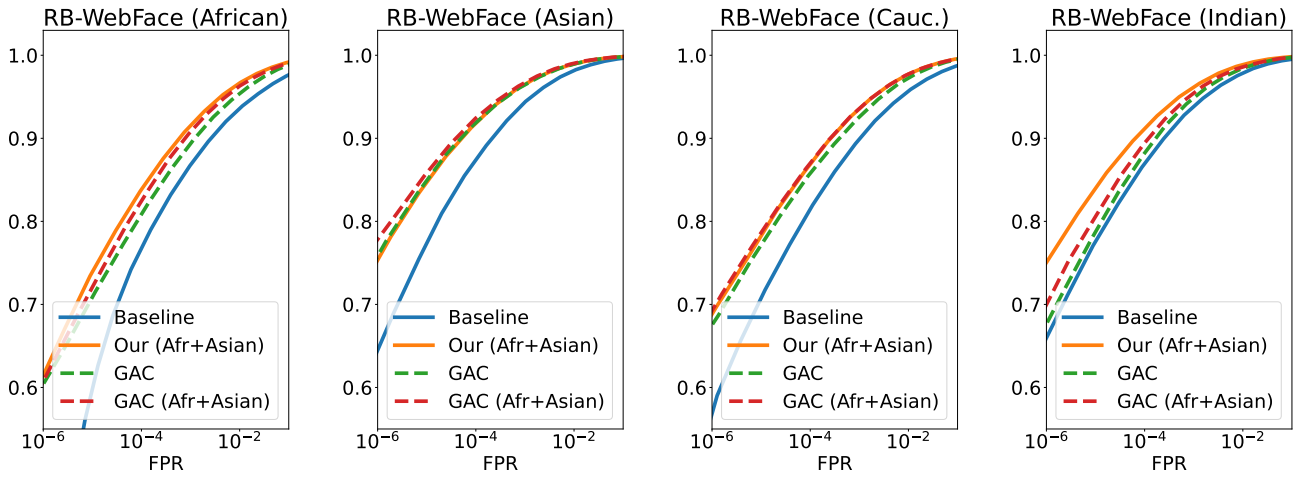
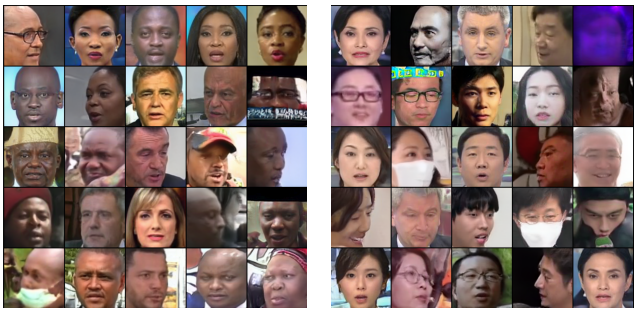


Figure 3: Comparison of the ROC curves for the methods on the newly assembled RB-WebFace validation set. Similarly to RFW, RB-WebFace consists of positive and negative pairs constructed from the set of samples. In the plot legend, *Ours* refers to our method, while *Baseline* stands for the *ArcFace R-50* baseline. *GAC* denotes *GAC (Ground truth)* method and *GAC (Afr+Asian)* describes its version with the proposed pretraining on $\mathcal{D}^{prior}$. Note the constant increase of TPR for the versions of algorithms enhanced by the proposed pretraining procedure.



(a) Generations for StyleGAN trained on AfricanFaceSet-5M

(b) Generations for StyleGAN trained on AsianFaceSet-3M

Figure 4: Random generations by a StyleGAN trained on either AfricanFaceSet-5M or AsianFaceSet-3M.

sampled uniformly from $\mathcal{D}^{prior}$ in 112×112 resolution. Since the output of the generator is of 128×128 resolution, we bilinearly downscale the generator output $\hat{I}$ to 112×112 px before calculating the loss, which is equal to $\|I - \hat{I}\|_2 + 0.8 \cdot \text{LPIPS}(I, \hat{I})$. The network follows IR-SE-50 architecture, which is an improved version of ResNet-50 with squeeze-and-excitation modules [3]. The encoder is trained for 16 million 1-sample steps. The training was performed on either 3 or 4 GPUs (either NVIDIA RTX 2080 TI or NVIDIA RTX 3090) with a minibatch of 48 or 64, respectively (depending on the experiment).

For the final fine-tuning stage (training the network for the face recognition task), we used the face.evoLVe [8] library for high-performance face recognition training, which was significantly modified. Before training, we copy all weights of the encoder from the first layer through the map2style blocks, excluding the latter, into the backbone, and attach a randomly initialized output block (BatchNorm + Dropout + fully-connected + BatchNorm, as recommended in the implementations (e.g. [8])). Additionally, we introduce a dropout layer with 0.15 dropout rate after every convolutional layer. The network is trained for 100 epochs. For the first 3 epochs, we freeze all layers except the first convolutional layer and the output block, and after the 3^{rd} epoch we unfreeze all layers. The optimizer is SGD with momentum of 0.9, weight decay of $2 \cdot 10^{-3}$, and the initial learning rate of 0.03 which is decreased by 1.5 every 5 epochs. Despite the fact that the learning rate setting, its scheduler, the introduction of the dropout layers, and higher weight decay compared to the standard ArcFace pipeline were modified, we found empirically that it helps consistently reproduce the results and better prevent overfitting in a general setting. Augmentations include resizing to 128×128, random cropping a 112×112 region, and horizontal flipping with 50% probability.

The network was trained on 3-5 GPUs (either NVIDIA RTX 2080 TI or NVIDIA RTX 3090) with batch size varying from 300 to 900, depending on the experiment (no significant sensitivity to the batch size parameter in that range was observed).

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 3

[2] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 2016. 3

[3] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[4] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 3

[5] 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. 3

[6] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 3

[7] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *International Conference on Computer Vision (ICCV)*, pages 692–702, 2019. 3

[8] Qingzhong Wang, Pengfei Zhang, Haoyi Xiong, and Jian Zhao. Face.evolve: A high-performance face recognition library. *arXiv preprint arXiv:2107.08621*, 2021. 5

[9] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 1