Supplementary Material for: StyleGAN-Human: A Data-Centric Odyssey of Human Generation

Jianglin Fu[1*], Shikai Li[1*], Yuming Jiang[2], Kwan-Yee Lin[1], Chen Qian[1], Chen Change Loy[2], Wayne Wu[1,3] ✉, and Ziwei Liu[2]

[1] SenseTime Research [2] S-Lab, Nanyang Technological University
[3] Shanghai AI Laboratory

1 SHHQ: StyleGAN-Human Datasets

Fig. 1: **Examples of raw data in the training dataset.**

The dataset we collected consists of $230K$ high-quality images of humans that vary in clothing appearance, ethnicity, and pose. Several training samples are shown in Figure 1. Please note that all these images are unprocessed. As shown in Figure 5, we also conduct qualitative comparisons with other human datasets to demonstrate the superiority of our clean, high-quality data. An additional comparison between the pruned DeepFashion and SHHQ is provided in Figure 2. The figure depicts the distribution statistics relating to age, body orientation,

* Equal contribution.
✉ Corresponding author (wuwenyan0503@gmail.com).

and face yaw angles. SHHQ achieves similar distributions of all these three attributes as DeepFashion, while the age groups of 10s and 40s are expanded more. Besides, SHHQ possesses a smoother body-orientation distribution. Besides, we display more generated human images from the baseline model trained with our SHHQ in Figure 6.

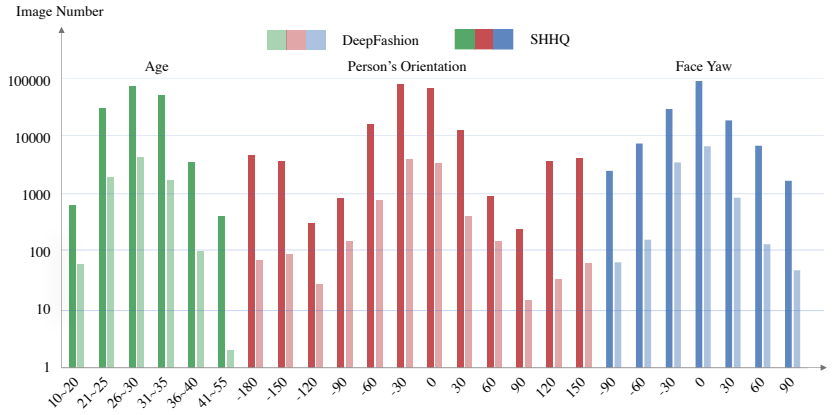

Fig. 2: **Extra data attributes comparison.** We show the difference between the pruned DeepFashion and our SHHQ on three attributes: age, person orientation, and face yaw angle. The units for the labels of "person's orientation" and "face yaw" are angles in degree.

2 Experiment Results

Table 1 and Figure 3 display the results of data size experiment. The results align with our expectation that increment training data will improve IS scores and reduce FID scores. Figure 7 and 8 depict the comparison between cropped faces and textures generated by the long-tail and uniform experiments. Due to privacy concerns, cropped training faces are not shown.

3 Training Scheme

We adopt the official NVIDIA Pytorch version of StyleGAN2-ADA as our codebase, and use the architecture of StyleGAN2. Here are several settings we use to accommodate this human generation task: (a) The input human-image has a width-to-height ratio of $1:2$, and the input resolution in the script is changed accordingly. (b) We adopt the same eight mapping layers as the original StyleGAN [3]. (c) There is no such a pretrained model for human images, so all the experiments are trained from scratch with the corresponding subset. (d) All other training hyper-parameters adopt the default values.

Table 1: **FID and IS for experiments of data size.** Quantitative comparisons at resolutions of 512×256 and 1024×512.

	Data Size	512×256 FID	IS	1024×512 FID	IS
$S0$	$10K$	7.80	3.87	7.23	3.93
$S1$	$20K$	4.46	4.40	4.33	4.56
$S2$	$40K$	2.61	4.81	2.80	4.92
$S3$	$80K$	2.53	4.90	2.09	5.01
$S4$	$160K$	2.09	4.92	2.02	5.04
$S5$	$230K$	1.97	5.04	1.57	5.02

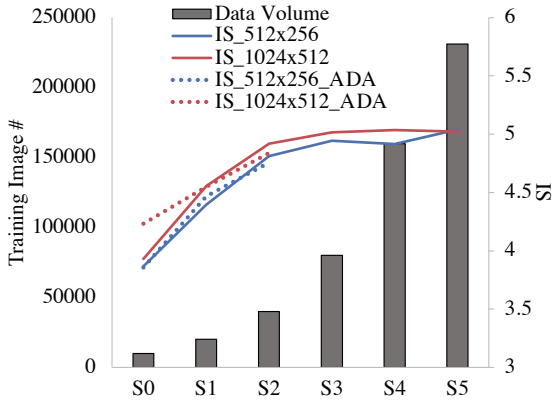

Fig. 3: **IS scores.** IS scores for experiments $S0$ - $S5$ in 1024×512 and 512×256 resolutions. Dotted lines represents the IS scores with ADA strategies.

4 Training strategy for data distribution experiments

The design of our distribution experiments focused more on whether the tail data could be improved under the premise of balancing different bins of the distribution. The other option could be implementing a customized data loader to sample outliers more frequently. Although the entire dataset can be fully leveraged in this way, the core of this strategy is more about augmenting distinct images in head data to improve general generation quality rather than emphasizing the tail part. Therefore we decide to sub-sample images with different distributions from the whole dataset.

5 Evaluation on other generative models

So far, there are no prior arts (either models or public datasets) designed specifically for the task of unconditional human generation. A considerable gap still

Fig. 4: **Failure cases from baseline model.** (a) - (c): Features such as face, texture and accessories are entangled. (d) Three hands detected on a single person. (e) - (f): Inferior generated hand quality. (g): Face quality could be better.

exists in generating high-quality humans, even under standard unposed input and output scenarios. Our work can be seen as a preliminary step toward the exploration of this task. In this work, we choose the StyleGAN family as the target model due to their promising baseline results and relatively complete ecosystem for downstream tasks. To further verify our experimental findings could be generalized to other models, we ablate alignment experiment on diffusion models [4]. The FID scores for the face-aligned, pelvis-aligned, and mid-body-aligned experiments are 4.29, 3.49, and 3.53, respectively. We use the same training images for these three experiments as the data-alignment experiments mentioned in the main paper. The model trained with face-aligned images produces the least satisfactory result, which is in line with our observation from StyleGAN training. For the other two settings, the pelvis-aligned experiment provides a slightly better FID (0.04) than the mid-body setting.

6 Limitations

Compared to face generation, training an unconditional human GAN is an arduous task because the semantic features of the full-body are much more complicated than a single face. Figure 4 shows some failure cases generated by the baseline model, suggesting several directions that can be strengthened in future human generation work. Artifacts caused by entangled features of faces/hands and clothing accessories are revealed in Figures 4 (a) - (c). Case (d) exhibits three hands on a person, which indicates that the global perception of the model needs to be improved [2]. We observe inferior hand quality in rare poses such as (e) and (f). To address this, the potential work could be augmenting training with such extreme poses, changing the data distribution, or implementing independent networks (i.e. fine-grained discriminators) to enhance local details [1]. The face and texture quality in cases (b) and (g) could be enhanced by local refinement as well.

7 Visualization of the Applications

7.1 Style-Mixing

Here we provide more examples of images generated by style-mixing on our baseline model. Figure 9, 10, and 11 represent the results of style-mixing on coarse, middle, and high resolution respectively. It shows that the latent at different scales control different high-level attributes of the clothed human, which is similar to face images.

7.2 Human Editing

Figure 12 displays the rotation of the human from the front view to the back view. The editing is done in W space. Figure 13 demonstrates the editing results in the length of sleeves and bottoms, based on StyleSpace [5].

References

1. Gadde, R., Feng, Q., Martinez, A.M.: Detail me more: Improving GAN's photo-realism of complex scenes. In: ICCV (2021) 4
2. Huang, R., Zhang, S., Li, T., He, R.: Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In: ICCV (2017) 4
3. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019) 2
4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021) 4
5. Wu, Z., Lischinski, D., Shechtman, E.: StyleSpace analysis: Disentangled controls for StyleGAN image generation. In: CVPR (2021) 5

Fig. 5: **Samples from different dataset with diverse resolution.**

Fig. 6: **Samples from our baseline model.** The model has shown the ability of generating random person with diverse clothing types, poses, genders, races, and hair types.

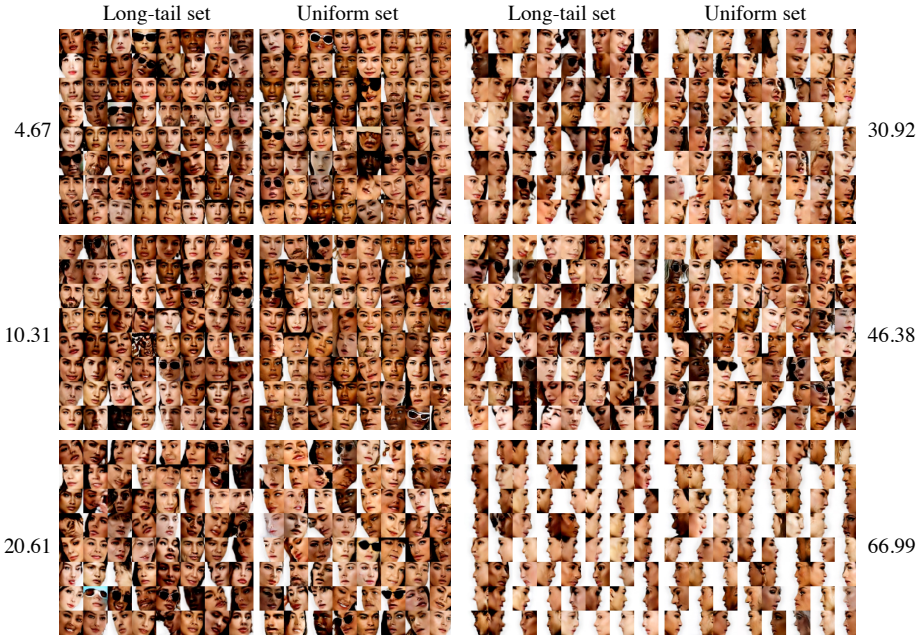

Fig. 7: **Cropped faces with different face yaw angles from each bin.** All the images are generated from the long-tail and uniform experiments.

Fig. 8: **Random cropped texture patches from each bin for both long-tail and uniform experiments.**

Fig. 9: **Style-mixing** with copying styles of *coarse* resolutions from reference images (top row), and rest spatial information are used from source images (first column).

Fig. 10: **Style-mixing** with copying styles of *middle* resolutions from reference images (top row), and rest spatial information are used from source images (first column).

Fig. 11: **Style-mixing** with copying styles of *fine* resolutions from reference images (top row), and rest spatial information are used from source images (first column).

Fig. 12: **Human Editing on orientation.**

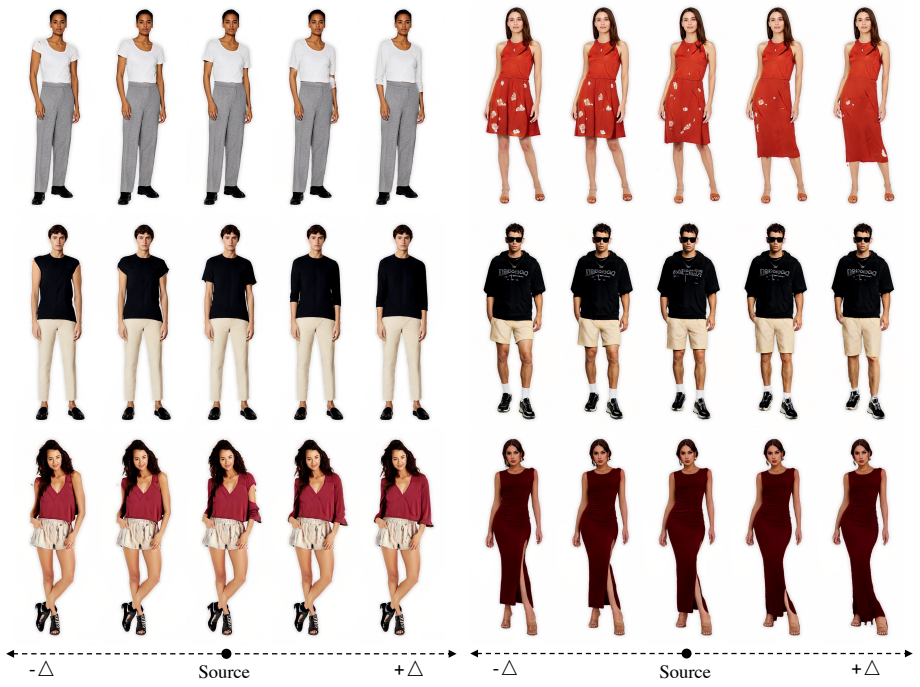

Fig. 13: **Human Editing on human sleeve length (left) and bottom length (right).**