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, Beijing, China
wuwenyan0503@gmail.com
[2] S-Lab, Nanyang Technological University, Singapore, Singapore
[3] Shanghai AI Laboratory, Shanghai, China

Abstract. Unconditional human image generation is an important task in vision and graphics, enabling various applications in the creative industry. Existing studies in this field mainly focus on "network engineering" such as designing new components and objective functions. This work takes a data-centric perspective and investigates multiple critical aspects in "data engineering", which we believe would complement the current practice. To facilitate a comprehensive study, we collect and annotate a large-scale human image dataset with over $230K$ samples capturing diverse poses and textures. Equipped with this large dataset, we rigorously investigate three essential factors in data engineering for StyleGAN-based human generation, namely data size, data distribution, and data alignment. Extensive experiments reveal several valuable observations *w.r.t.* these aspects: 1) Large-scale data, more than $40K$ images, are needed to train a high-fidelity unconditional human generation model with a vanilla StyleGAN. 2) A balanced training set helps improve the generation quality with rare face poses compared to the long-tailed counterpart, whereas simply balancing the clothing texture distribution does not effectively bring an improvement. 3) Human GAN models that employ body centers for alignment outperform models trained using face centers or pelvis points as alignment anchors. In addition, a model zoo and human editing applications are demonstrated to facilitate future research in the community. Code and models are publicly available (Project page: https://stylegan-human.github.io/. Code and models: https://github.com/stylegan-human/StyleGAN-Human.)

Keywords: Human image generation · Data-centric · StyleGAN

J. Fu and S. Li—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19787-1_1.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13676, pp. 1–19, 2022.
https://doi.org/10.1007/978-3-031-19787-1_1

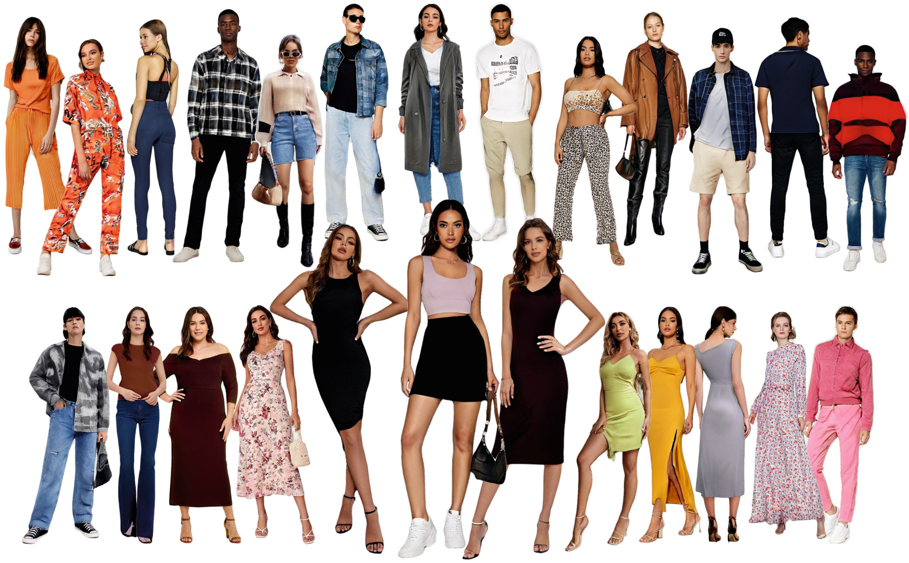

Fig. 1. A data-centric odyssey of human generation. With good "data engineering" practices, the StyleGAN-Human model could generate high-resolution photo-realistic human images as presented. Zoom in for the best view.

1 Introduction

Generating photo-realistic images of clothed humans unconditionally can provide great support for downstream tasks such as human motion transfer [8,44], digital human animation [42], fashion recommendation [32,40], and virtual try-on [15,56,85]. Traditional methods create dressed humans with classical graphics modeling and rendering processes [17,31,50,61,63,70,78,93]. Although impressive results have been achieved, these prior works are easy to suffer from the limitation of robustness and generalizability in complex environments. Recent years, Generative Adversarial Networks (GANs) have demonstrated remarkable abilities in real-world scenarios, generating diverse and realistic images by learning from large-quantity and high-quality datasets. [24,33,36,66].

Among the GAN family, StyleGAN2 [37] stands out in generating faces and simple objects with unprecedented image quality. A major driver behind recent advancements [2,34,37,80,90] on such StyleGAN architectures is the prosperous discovery of "network engineering" like designing new components [2,34,90] and loss functions [37,80]. While these approaches show compelling results in generating diverse objects (*e.g.*, faces of humans and animals), applying them to the photo-realistic generation of articulated humans in natural clothing is still a challenging and open problem.

In this work, we focus on the task of Unconditional Human Generation, with a specific aim to train a good StyleGAN-based model for articulated humans from a *data-centric* perspective. First, to support the data-centric investigation, collecting a large-scale, high-quality, and diverse dataset of human bodies in clothing is necessary. We propose the Stylish-Humans-HQ Dataset (SHHQ), which contains $230K$ clean full-body images with a resolution of 1024×512 at least and up to 2240×1920. The SHHQ dataset lays the foundation for extensive experiments on unconditional human generation. Second, based on the proposed SHHQ dataset, we investigate three fundamental and critical questions that were not thoroughly discussed in prior works and attempt to provide useful insights for future research on unconditional human generation.

To extract the questions that are indeed *important* for the community of Unconditional Human Generation, we make an extensive survey on recent literature in the field of general unconditional generation [5,6,20,24,33,36,51]. Based on the survey, three questions that are investigated actively can be concluded as below. **Question-1**: What is the relationship between the *data size* and the generation quality? Several previous works [6,27,34,81,98] pointed out that the quantity of training data is the primary factor to determine the strategy for improving image quality in face and other object generation tasks. In this study, we want to examine the minimum quantity of training data required to generate human images of high quality without any extensive "network engineering" effort. **Question-2**: What is the relationship between the *data distribution* and the generation quality? This question has received extensive attention [14,22,52,71,92] and leads to a research topic dealing with data imbalance [46]. In this study, we aim to exploit data imbalance problem in the human generation task. **Question-3**: What is the relationship between the scheme of *data alignment* and the generation quality? Different alignment schemes applied to uncurated faces [36,38] and non-rigid objects [6,13,72] show success in enhancing training performance. In this study, we seek a better data alignment strategy for human generation.

Based on the proposed SHHQ dataset and observations from our experiments, we establish a Model Zoo with three widely-adopted unconditional generation models, *i.e.*, StyleGAN [36], StyleGAN2 [37], and alias-free StyleGAN [35], in both resolution of 1024×512 and 512×256. Although hundreds of StyleGAN-based studies exist for *face* generation/editing tasks, a high-quality and public model zoo for *human* generation/editing with StyleGAN family is still missing. We believe the provided model zoo has great potentials in many human-centric tasks, *e.g.*, human editing, neural rendering, and virtual try-on.

We further construct a human editing benchmark by adapting previous editing methods based on facial models to human body models (*i.e.*, PTI [68] for image inversion, InterFaceGAN [75], StyleSpace [90], and SeFa [76] for image manipulation). The impressive results in editing human clothes and attributes demonstrate the potential of the given model zoo in downstream tasks. In addition, a concurrent work, InsetGAN [16], is evaluated with our baseline model, further showing the potential usage of our pre-trained generative models.

Here is the summary of the main contributions of this paper: **1)** We collect a large-scale, high-quality, and diverse dataset, Stylish-Humans-HQ (SHHQ), containing $230K$ human full-body images for unconditional human generation task. **2)** We investigate three crucial questions that have aroused broad interest in the community and discuss our observation through comprehensive analysis.

3) We build a model zoo for unconditional human generation to facilitate future research. An editing benchmark is also established to demonstrate the potential of the proposed model zoo.

2 Related Work

2.1 Dataset for Human Generation

Large-scale and high-quality clothed human-centric training datasets are the critical fuel for the training of StyleGAN models. A qualified dataset should conform to the following aspects: **1)** *Image quality*: high-resolution images with rich textures offer more raw detailed semantic information to the model. **2)** *Data volume*: the size of dataset should be sufficient to avoid generative overfitting [4, 97]. **3)** *Data coverage*: the dataset should cover multiple attribute dimensions to guarantee diversity of the model, for instance, gender, clothing type, clothing texture, and human pose. **4)** *Data content*: since this report only focuses on the generation of single full-body human, occlusion caused by other people or objects is not considered here, whereas self-occlusion is taken into account. That is, each image should contain only one complete human body.

Publicly available datasets built particularly for full human-body generation are rare, but there are several practices [30,48,49,77] cooperating with DeepFashion [45] and Market1501 [99]. DeepFashion dataset [30,45] with well-labeled attributes and diverse garment categories is satisfactory for image classification and attribute prediction, but not adequate for unconditional human generation since it emphasizes fashion items rather than human bodies. Thus the number of close-up shots of clothing is much higher than that of full-body images. Market1501 dataset [99] fails for human generation tasks due to its low resolution (128×64). There are some human-related datasets in other domains rather than GAN-based applications: datasets related to human parsing [19,43] are limited by scalability and diversity; common datasets for virtual try-on tasks either contain only the upper body [25] or are not public [96]. A detailed comparison of the above datasets in terms of data scale, average resolution, attributes labeling, and proportion of full-body images across the whole dataset is listed in Table 1. In general, there is no high-quality and large-scale full human-body dataset publicly available for the generative purpose.

2.2 StyleGAN

In recent years, the research focus has gradually shifted to generating high-fidelity and high-resolution images through Generative Adversarial Networks [6, 33]. The StyleGAN generator [36] was introduced and became the state-of-the-art network of unconditional image generation. Compared to previous GAN-based architectures [5,24,55], SytleGAN injects a separate attribute factor (i.e., style) into the generator to influence the appearance of generated images. Then StyleGAN2 [37] redesigns the normalization, multi-scale scheme, and regularization method to rectify the artifacts in StyleGAN images. The latest update to

Table 1. Comparison of SHHQ with other publicly available datasets.

Dataset	Total image #	Mean resolution	Labeled attributes	Full-body ratio
ATR [43]	7,700	400 × 600	✓	76%
Mark1et1501 [99]	32,668	128 × 64	✓	100%
DeepFashion [45]	146,680	1101 × 750	✓	6.8%
LIP [19]	50,462	196 × 345	✓	37%
VITON [25]	16,253	256 × 192	✗	0%
SHHQ	**231,176**	**1024 × 512**	✓	**100%**

StyleGAN [35] reveals the non-ideal case of detailed textures sticking to fixed pixel locations and proposes an alias-free network.

2.3 Image Editing

Benefiting from StyleGAN, one of the significant downstream applications is image editing [1,60,75,90,95]. A standard image editing pipeline usually involves inversion from a real image to the latent space and manipulating the embedded latent code. Existing works for *image inversion* can be categorized into optimization-based [2,79], encoder-based [82,86], and hybrid methods [68], which exploit encoders to embed images into latent space and then refine with optimization. As for *image manipulation*, studies explore the capability of attribute disentanglement in the latent space in supervised [29,75,90] or unsupervised [26,76,83] manners. In specific, Jiang *et al.* [29] proposes to use fine-grained annotations to find non-linear manipulation directions in the latent space, while SeFa [76] searchs for semantic directions without supervision. StyleSpace [90] defines the style space S and proves that it is more disentangled than W and $W+$ space.

3 Stylish-Humans-HQ Dataset

To investigate the key factors in unconditional human generation task from a data-centric perspective, we propose a large-scale, high-quality, and diverse dataset, Stylish-Humans-HQ (SHHQ). In this section, we first present the data collection and preprocessing (Sect. 3.1), in which we construct the SHHQ dataset. Then, we analyze the data statistic (Sect. 3.2) to demonstrate the superiority of SHHQ compared to other datasets from a statistical perspective.

3.1 Data Collection and Preprocessing

Over 500K raw data were collected legally in two ways: 1) *From the Internet.* We crawled images, with CC-BY-2.0 licenses available, mainly from Flickr, Unsplash, Pixabay and Pexels, by searching keywords related to humans. 2) *From data*

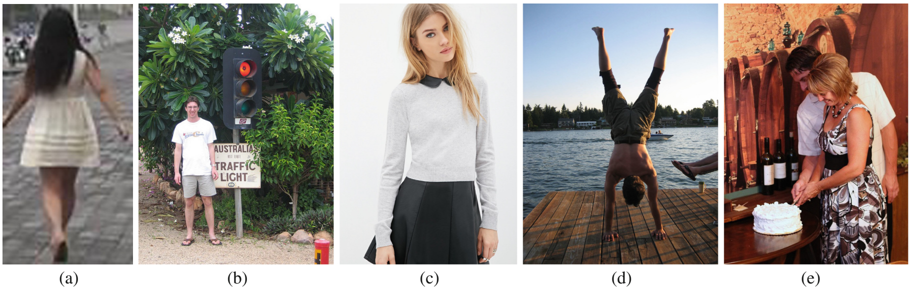

Fig. 2. Data preprocessing. The following types of images will be removed during our data preprocessing pipeline. (a) Low resolution. (b) Not placed in the center. (c) Missing body parts. (d) Extreme posture. (e) Multi-person.

providers. We purchased images from individual photographers, model agencies, and other providers' databases. Images were reviewed by our institute's legal team before the purchase, to ensure the permission of usage in research. We preprocess the data with six factors taken into consideration (*e.g.*, resolution [45], body position [36], body-part occlusion, human pose [36,45], multi-person, and background), which are critical for the quality of a human dataset. After the data preprocessing, we obtain a clean dataset of 231,176 images with high quality; see Fig. 5(a) for examples. We filter the images according to following aspects. **1) Resolution:** We discard images lower than 1024 × 512 resolution (Fig. 2(a)). **2) Body Position:** The position of the body varies widely in different images, *i.e.*, Fig. 2(b). We design a procedure in which each person is appropriately cropped based on human segmentation [12], padded and resized to the same scale, and then placed in the image such that the body center is aligned. The body center is defined as the average coordinate of the entire body using segmentation. **3) Body-Part Occlusion:** This work aims at generating full-body human images, images with any missing body parts are removed (*e.g.*, the half-body portrait shown in Fig. 2(c)). We remove images with extreme poses (*e.g.*, lying postures, handstand in Fig. 2(d)) to ensure learnability of the data distribution. We exploit human pose estimation [7] to detect those extreme poses. **4) Multi-Person Images:** Some images contain multiple persons, such as Fig. 2(e). The goal of this work is to generate single full-body person, so we keep unoccluded single-person full-body images, and remove those with occluded people. **5) Background:** Some images contain complicated backgrounds, requiring additional representation ability. To focus on the generation of the human body itself and eliminate the influence of various backgrounds, we use a segmentation mask [12] to modify the image background to pure white. The edges of the mask are then smoothed by Gaussian blur.

3.2 Data Statistics

Table 1 presents the comparison between SHHQ and other public datasets from the following three aspects: **1) Dataset Scale:** As shown in the table, our

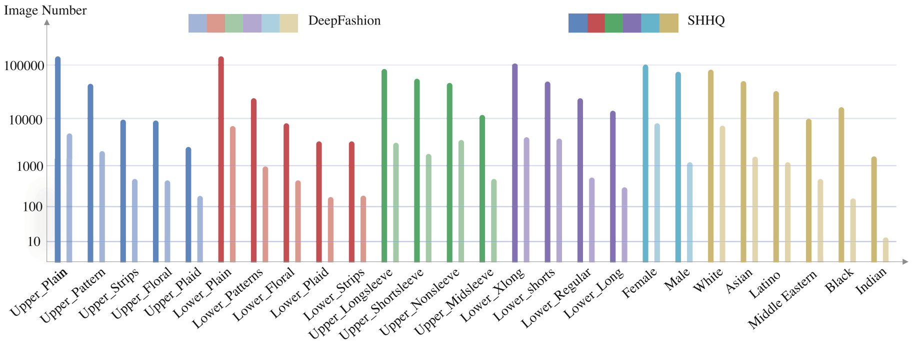

Fig. 3. Attribute distribution. Comparison of different attributes between the pruned DeepFashion and SHHQ dataset: Texture/length of the upper/lower clothing, gender, and ethnicity. (More attributes comparison in supplementary).

proposed SHHQ is currently the largest dataset in scale compared to others. Among them, the data volume of SHHQ is 1.6 times that of DeepFashion [45] dataset and is much larger than that of others. **Resolution.** Images from ATR [43], Market1501 [99], LIP [19], and VITON [25] are lower in resolution, which is insufficient for our generation task, while the proposed SHHQ and DeepFashion provide high-definition images up to 2240 × 1920. **2) Labels:** All datasets beside VITON provide various labeled attributes. Specifically, DeepFashion [30,45] and SHHQ label the clothing types and textures, which is useful for human generation/editing tasks. **3) Full-Body Ratio:** This number denotes the proportion of full-body images in the dataset. Although DeepFashion [45] offers over $146K$ images with decent resolution, only 6.8% of them are full-body images, while SHHQ achieves a 100% full-body ratio. The visual comparison among these datasets and the proposed SHHQ dataset is shown in supplementary.

In sum, SHHQ covers the largest number of human images with high-resolution, labeled clothing attributes, and 100% full-body ratio. It again confirms that SHHQ is more suitable for full-body human generation than other public datasets.

Of all the datasets compared above, DeepFashion [45] is the most relevant to our human generation task. In Fig. 3, we further present the comparison of different attributes between filtered DeepFashion [45] (full-body only) and SHHQ in a more detailed view. The bar chart depicts the distributions along six dimensions: upper cloth texture, lower cloth texture, upper cloth length, lower cloth length, gender, and ethnicity. In particular, the number of females is approximately 4 times the number of males in filtered DeepFashion [45], while our dataset features a more balanced female-to-male ratio of 1.49. With the help of DeepFace API [74], it is shown that SHHQ is more diverse in terms of ethnicity. Advantages are also shown in the other five attributes. In terms of garment-related attributes, images with specific labels in filtered DeepFashion [45] are too scarce to be used as a training set. The Stylish-Humans-HQ dataset boosts the number of each category by an average of 24.4 times.

4 Systematic Investigation

We conduct extensive experiments to study three factors concerning the quality of generated images: 1) data size (Sect. 4.1), 2) data distribution (Sect. 4.2), and 3) data alignment (Sect. 4.3). Our investigations are all built on the StyleGAN2 architecture and codebase. More implementation details and experimental results can be found in supplementary.

4.1 Data Size

Motivation. Data size is an essential factor that determines the quality of generated images. Previous literature takes different strategies to improve the generation performance according to different dataset sizes: regularization techniques [6] are employed to train a large dataset, while augmentation [34,81,98] and conditional feature transferring [52,88] are proposed to tackle the limited data. Here, we design sets of experiments to examine the relationship between training data size and the image quality of generated humans.

Experimental Settings. To determine the relationship between data size and image quality for the unconditional human GAN, we construct 6 sub-datasets and denoted these subsets as $S0$ ($10K$), $S1$ ($20K$), $S2$ ($40K$), $S3$ ($80K$), $S4$ ($160K$) and $S5$ ($230K$). Here, $S0$ is the pruned DeepFashion dataset. We perform the training on two resolution settings for each set: 1024×512 and 512×256. Considering the case of limited data, we also conduct additional training experiments with adaptive discriminator augmentation (ADA) [34] for small datasets $S0$, $S1$, and $S2$. Fréchet Inception Distance (FID) and Inception Score (IS) are the indicators for evaluating the model performance.

Results. As shown in Fig. 4(a), the FID scores (solid lines) decrease as the size of the training dataset increases for both resolution settings. The declining trend is gradually flattening and tends to converge. $S0$ generates the least satisfactory results, with FID of 7.80 and 7.23 for low- and high-resolution, respectively, while $S1$ achieves corresponding improvements of 42% and 40% on FID with only an additional $10K$ training images. When the training size reaches $40K$ for both resolutions, the FID curves start to converge to a certain extent. The dotted lines indicate the results of ADA experiments with subsets $S0$–$S2$. The employed data augmentation strategy helps to reduce FID when training data is less than $40K$. More quantitative results (FID/IS) are in supplementary.

Discussion. The experiments confirm that ADA can improve the generation quality for datasets smaller than $40K$ images, in terms of FID and IS. However, ADA still cannot fully compensate for the impact of insufficient data. Besides, when the amount of data is less than $40K$, the relationship between image quality and data size is close to linear. As the amount of data increases to $40K$ and more, the improvement in the image quality slows down and is less significant.

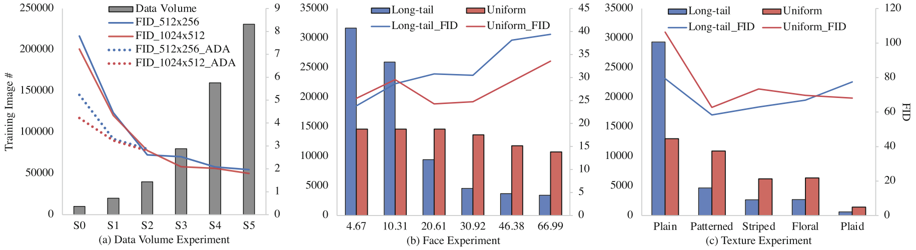

Fig. 4. Experiment results. (a) FIDs for experiments $S0$–$S5$ in 1024×512 and 512×256 resolutions. Dotted lines shows the FIDs of the models trained with ADA. (b) Bin-wise FIDs of long-tailed and uniform distribution in terms of facial yaw angle along with the number of training images. (c) Bin-wise texture FIDs of long-tailed/uniform distribution along with the number of training images. (Color figure online)

4.2 Data Distribution

Motivation. The nature of GAN makes the model inherits the distribution of the training dataset and introduces generation bias due to dataset imbalance [46]. This bias severely affects the performance of GAN models. To address this issue, studies for unfairness mitigation [14,22,52,71,92] have attracted substantial research interest. In this work, we explore the question of data distribution in human generation and conduct experiments to verify whether a uniform data distribution can improve the performance of a human generation model.

Experimental Settings. This study decomposes the distribution of the human body into Face Orientation and Clothing Texture, since face fidelity has a significant impact on visual perception and clothing occupies a large portion of the full-body image. The general features of human faces are relatively symmetrical; thus, we fold yaw distribution vertically along 0° and get the long-tailed distribution. For the face and clothing experiments, we collect an equal number of long-tailed and uniformly distributed datasets from SHHQ for face rotation angle and upper-body clothing texture, respectively.

Results. To evaluate the image quality in terms of different distributions, the cropped faces and clothing regions are used to calculate FID, and FID is calculated separately for each bin. Result can be found in Fig. 4(b) and (c).

1) Face Orientation: As for the long-tailed experiment (blue curve in Fig. 4(b)), the FID progressively grows as the face yaw angle increases and remains high when the facial rotation angle is too large. By contrast, the upward trend for the face FID in the uniform experiment (red) is more gradual. In addition, the amount of the training data of the first two bins in the uniform set is greatly reduced compared to the long-tail experiment, but the damage to FID is slight.
2) Clothing Texture: From Fig. 4(c), except for the first bin ("plain" pattern), the FID curve climbs steadily as the amount of training data for the long-tailed experiment decreases, and the FID curve for the uniform experiment

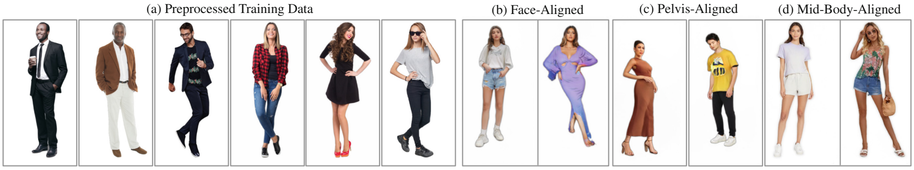

Fig. 5. Example of preprocessed data and different alignment schemes. Part (a) shows processed training data with the consideration of resolution, body position, body-part occlusion, human pose, multi-person and background. (b)–(d) display random samples with three different alignment strategies.

also shows a near-uniform pattern. In particular, FID of the last bin for the uniform experiment is lower than that in the long-tailed setting. We infer that the training samples for "plaid" clothing texture in the long-tailed experiment are too few to be learned by the model. As for the "plain" bin results, the long-tailed distribution has a lower FID score in this bin. The reason may lie in that the number of plain textures in the long-tailed distribution is considerably higher than that in the uniform distribution. Also, it can be observed that the training patches in this bin are mainly textureless color blocks where such patterns may be easier to capture by models.

Discussion. Based on the above analysis, we conclude that the uniform distribution of face rotation angles can effectively reduce the FID of rare training faces while maintaining acceptable image quality for the dominant faces. However, simply balancing the distribution of texture patterns does not always reduce the corresponding FID effectively. This phenomenon raises an interesting question that can be further explored: is the relation between image quality and data distribution also entangled with other factors, *e.g.*, image pattern and data size? Additionally, due to the nature of GAN-based structures, a GAN model memorizes the entire dataset, and usually, the discriminator tends to overfit those poorly sampled images at the tail of the distribution. Consequently, the long-tailed situation accumulated as "tail" images is barely generated. From this perspective, it also can be seen that the uniform distribution preserves the diversity of faces/textures and partially alleviates this problem.

4.3 Data Alignment

Motivation. Recently, researchers have drawn attention to spatial bias in generation tasks. Several works [36,38] align face images with keypoints for face generation, and other studies propose different alignment schemes to preprocess non-rigid objects [6,13,28,47,72]. In this paper, we study the relationship between the spatial deviation of the entire human and the generated image quality.

Experimental Settings. We randomly sample a set of $50K$ images from the SHHQ dataset and align every image separately using three different alignment strategies: aligning the image based on the face center, pelvis, and the midpoint of the whole body, as shown in Fig. 5.

Following are the reasons for selecting these three positions as alignment centers. 1) For the face center, we hypothesize that faces contain rich semantic information that is valuable for learning and may account for a heavy proportion in human generation. 2) For the pelvis, studies related to human pose estimation [53,57,62,84] conventionally predict the body joint coordinates relative to the pelvis. Thus we employ the pelvis as the alignment anchor. 3) For the body's midpoint, the leg-to-body ratio (the proportion of upper and lower body length) may vary among different people; therefore, we try to find the mean coordinates of the full body with the help of the segmentation mask.

Results. Human images are complex and easily affected by various extrinsic factors such as body poses and camera viewpoints. The FID scores for the face-aligned, pelvis-aligned, and mid-body-aligned experiments are 3.5, 2.8, and 2.4, respectively. Figure 5 further interprets this perspective as the human bodies in (b) and (c) are tilted, and the overall image quality is degraded. The example shown in Fig. 5(c) also presents the inconsistent human positions caused by different leg-to-body ratios.

Discussion. Both FID scores and visualizations suggest that the human generative models gain more stable spatial semantic information through the mid-body alignment method than face- and pelvis-centered methods. We believe this observation could benefit later studies on human generation.

4.4 Experimental Insights

Now the questions can be answered based on the above investigations:

For **Question-1** (Data Size): A large dataset with more than $40K$ images helps to train a high-fidelity unconditional human generation model, for both 512×256 and 1024×512 resolution.

For **Question-2** (Data Distribution): The uniform distribution of face rotation angles helps reduce the FID of rare faces while maintaining a reasonable quality of dominant faces. But simply balancing the clothing texture distribution does not effectively improve the generation quality.

For **Question-3** (Data Alignment): Aligning the human by the center of the full body presents a quality improvement over aligning the human by face or pelvis centers.

5 Model Zoo and Editing Benchmark

5.1 Model Zoo

In the field of face generation, a pre-trained StyleGAN [36] model has shown remarkable potential and success in various downstream tasks, including editing [2,90], neural rendering [23], and super-resolution [11,54]. Nevertheless, a

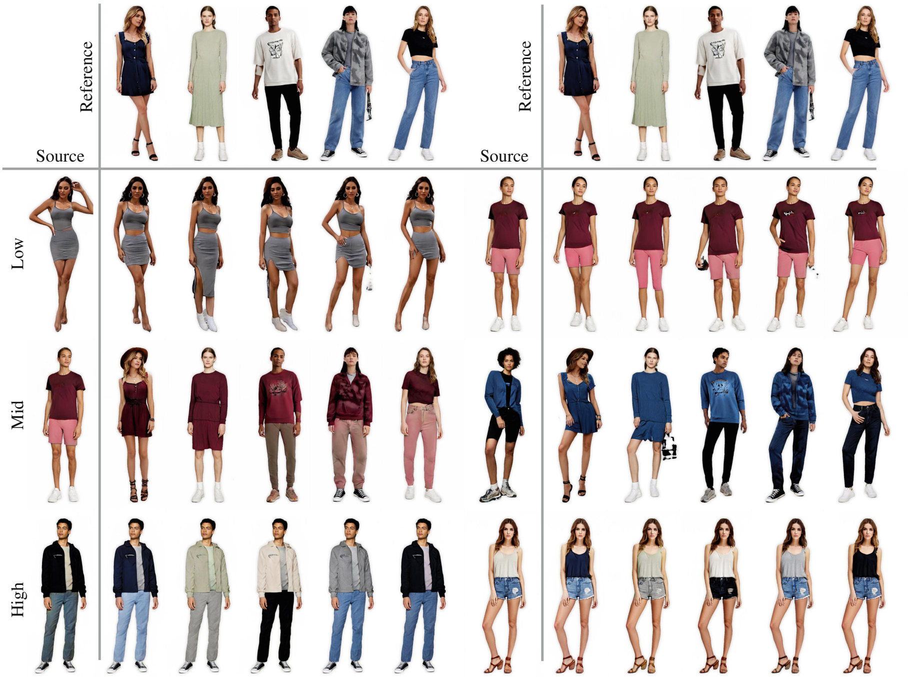

Fig. 6. Style-mixing results. The reference and source images are randomly sampled from the provided baseline model. The rest of the images are generated by style-mixing: borrowing low/mid/high layers in the reference images' latent codes and combining them with the rest layers of latent code in source images.

publicly available pre-trained model is still lacking for the human generation task. To fill this gap, we train our baseline model on SHHQ using the StyleGAN2 [37] model, which provides the best FID of 1.57. As seen in Figs. 1, our model has the ability to generate full-body images with diverse poses and clothing textures under satisfactory image quality. To adapt various application scenarios, we build a model zoo consisting of trained models from different StyleGAN architectures [35–37] in both resolution (1024 × 512 and 512 × 256).

Furthermore, the style mixing results of the baseline model show the interpretability of the corresponding latent space. As seen in the Fig. 6, source and reference images are sampled from the baseline model, and the rest images are the style-mixing results. We see that copying low layers from reference images to source images brings changes in geometry features (pose) while other features such as skin, garments, and identities in source images are preserved. When replicating middle styles, the source person's clothing type and identical appearance are replaced by reference. Finally, we observe that fine styles from high-resolution layers control the clothing color. These results suggest that the provided model's geometry and appearance information are well disentangled.

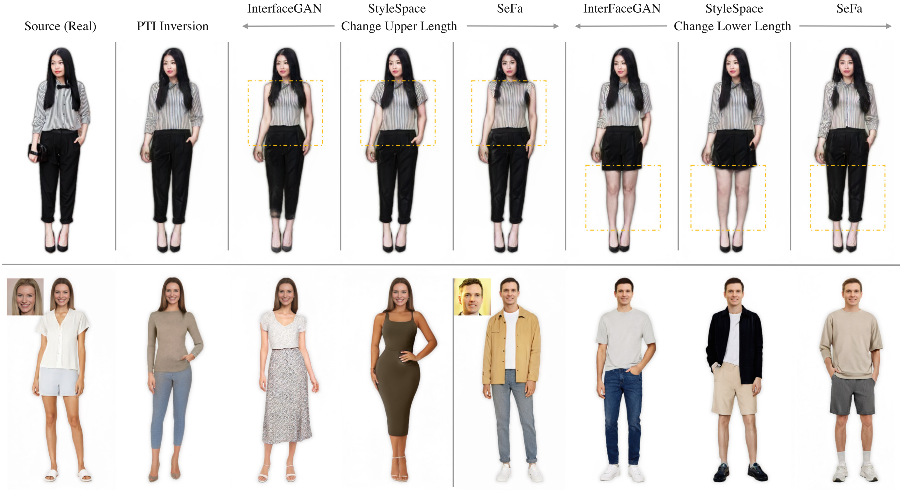

Fig. 7. Image editing and InsetGAN results. *Top row* presents editing results on an real image (left) after PTI inversion. The length of sleeve and pants are edited using different techniques. *Bottom Row* shows the InsetGAN results of different human bodies generated from the given baseline model and two faces generated from the FFHQ [37] model.

5.2 Editing Benchmark

StyleGAN has presented remarkable editing capabilities over faces. We extend it to the full-scale human by using off-the-shelf inversion and editing methods, in which we validate the potential of our proposed model zoo. We also re-implement the concurrent human generation method, InsetGAN [16], to further demonstrate another practical usage.

First, we leverage several SOTA StyleGAN-based facial editing techniques, such as InterFaceGAN [75], StyleSpace [90], and SeFa [76], with multiple editing directions: garment length for tops and bottoms, and global pose orientation. To examine the ability of editing real images with the provided model, we trained the e4e encoder [82] on SHHQ to obtain the inverted latent code as initial pivot. PTI [68] is then used to fine-tune the generator for each specific image.

As illustrated in Fig. 7, PTI presents the ability to invert real full-body human images. For attributes manipulation, StyleSpace [90] expresses better disentanglement compared to InterFaceGAN [75] and SeFa [76], as only the attribute-targeted region has been changed. However, as for the regions to be edited, the results of InterFaceGAN [75] are more natural and photo-realistic. It turns out that the latent space of the human body is more complicated than other domains such as faces, objects, and scenes, and more attention should be paid to disentangle human attributes. More editing results are shown in supplementary.

We re-implement InsetGAN [16] by iteratively optimizing the latent codes for random faces and bodies generated by the FFHQ [37] and our model, respectively. In the bottom row of Fig. 7, we show the fused full-body images with different male and female faces. The optimization procedure blends diverse faces and bodies in a graceful manner. Both the adopted editing methods and the multi-GAN optimization method demonstrate the effectiveness and convenience of our provided model zoo and verify its potential in human-centric tasks.

6 Future Work

In this study, we take a preliminary step towards the exploration of the human generation/editing tasks. We believe many future works can be further explored based on the SHHQ dataset and the provided model zoo. In the following, we discuss three interesting directions, *i.e.*, Human Generation/Editing, Neural Rendering, and Multi-modal Generation.

Human Generation/Editing. Studies in human generation [16,96], human editing [3,21,69], virtual try-on [15,41,56,85], and motion transfer [8,44] heavily rely on large datasets to train or use existing pre-trained models as the first step of transfer learning. Furthermore, editing benchmarks show that disentangled editing of the human body remains challenging for existing methods [75,90]. In this context, the released model zoo could expedite such research progress. Additionally, we further analyze failure cases generated by the provided model and discuss corresponding potential efforts that could be made to human generation tasks in supplementary.

Neural Rendering. Another future research direction is to improve 3D consistency and mitigate artifacts in full-body human generation through neural rendering [9,10,23,58,59,73]. Similar to work such as EG3D [9], StyleNeRF [23], and StyleSDF [59], we encourage researchers to use our human models to facilitate human generation with multi-view consistency.

Multi-modal Generation. Cross-modal representation is an emerging research trend, such as CLIP [65] and ImageBERT [64]. Hundreds of studies are made on text-driven image generation and manipulation [29,30,39,60,64,67,87,91,94], *e.g.*, DALLE [67] and AttnGAN [94]. In the meantime, several studies show interest in probing the transfer learning benefits of large-scale pre-trained models [18,65,89]. Most of these works focus on faces and objects, whereas research fields related to full-scale humans could be explored more, *e.g.*, text-driven human attributes manipulation, with the help of the provided full-body human models.

7 Conclusion

This work mainly probes how to train unconditional human-based GAN models to generate photo-realistic images from a data-centric perspective. By leveraging the 230K SHHQ dataset, we analyze three fundamental yet critical issues that the community cares most about: data size, data distribution, and data alignment. While experimenting with StyleGAN and large-scale data, we obtain several empirical insights. Apart from these, we create a model zoo, consisting of six human-GAN models, and the effectiveness of the model zoo is demonstrated by employing several state-of-the-art face editing methods.

Acknowledgements. This work is supported by NTU NAP, MOE AcRF Tier 1 (2021-T1-001-088), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2StyleGAN++: how to edit the embedded images? In: CVPR (2020)
2. Abdal, R., Zhu, P., Mitra, N.J., Wonka, P.: StyleFlow: attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. ACM TOG **40**(3), 1–21 (2021)
3. Albahar, B., Lu, J., Yang, J., Shu, Z., Shechtman, E., Huang, J.B.: Pose with style: detail-preserving pose-guided image synthesis with conditional StyleGAN. ACM TOG **40**(6), 1–11 (2021)
4. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. In: ICLR (2017)
5. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: ICML (2017)
6. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2019)
7. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR (2017)
8. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: ICCV (2019)
9. Chan, E.R., et al.: Efficient geometry-aware 3D generative adversarial networks. In: CVPR (2022)
10. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: Pi-GAN: periodic implicit generative adversarial networks for 3D-aware image synthesis. In: CVPR (2021)
11. Chan, K.C., Wang, X., Xu, X., Gu, J., Loy, C.C.: GLEAN: generative latent bank for large-factor image super-resolution. In: CVPR (2021)
12. MMSegmentation Contributors: MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark (2020). https://github.com/open-mmlab/mmsegmentation
13. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: NeurIPS (2021)
14. Dionelis, N., Yaghoobi, M., Tsaftaris, S.A.: Tail of distribution GAN (TailGAN): GenerativeAdversarial-network-based boundary formation. In: SSPD (2020)
15. Dong, H., et al.: Towards multi-pose guided virtual try-on network. In: ICCV (2019)
16. Frühstück, A., Singh, K.K., Shechtman, E., Mitra, N.J., Wonka, P., Lu, J.: InsetGAN for full-body image generation. In: CVPR (2022)

17. Gahan, A.: 3ds Max Modeling for Games: Insider's Guide to Game Character, Vehicle, and Environment Modeling (2012)
18. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: CVPR (2019)
19. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In: CVPR (2017)
20. Goodfellow, I., et al.: Generative adversarial nets. In: NeurIPS (2014)
21. Grigorev, A., et al.: StylePeople: a generative model of fullbody human avatars. In: CVPR (2021)
22. Grover, A., et al.: Bias correction of learned generative models using likelihood-free importance weighting. In: NeurIPS (2019)
23. Gu, J., Liu, L., Wang, P., Theobalt, C.: StyleNeRF: a style-based 3d aware generator for high-resolution image synthesis. In: ICLR (2022)
24. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein GANs. In: NeurIPS (2017)
25. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: VITON: an image-based virtual try-on network. In: CVPR (2018)
26. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: GanSpace: discovering interpretable GAN controls. In: NeurIPS (2020)
27. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Deceive D: adaptive pseudo augmentation for GAN training with limited data. In: NeurIPS (2021)
28. Jiang, Y., Chan, K.C., Wang, X., Loy, C.C., Liu, Z.: Robust reference-based super-resolution via C2-matching. In: CVPR (2021)
29. Jiang, Y., Huang, Z., Pan, X., Loy, C.C., Liu, Z.: Talk-to-edit: fine-grained facial editing via dialog. In: ICCV (2021)
30. Jiang, Y., Yang, S., Qiu, H., Wu, W., Loy, C.C., Liu, Z.: Text2Human: text-driven controllable human image generation. ACM TOG **41**(4), 1–11 (2022)
31. Jojic, N., Gu, J., Shen, T., Huang, T.S.: Computer modeling, analysis, and synthesis of dressed humans. TCSVT **9**(2), 378–388 (1999)
32. Kang, W.C., Fang, C., Wang, Z., McAuley, J.: Visually-aware fashion recommendation and design with generative image models. In: ICDM (2017)
33. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: ICLR (2017)
34. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: NeurIPS (2020)
35. Karras, T., et al.: Alias-free generative adversarial networks. In: NeurIPS (2021)
36. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
37. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: CVPR (2020)
38. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: CVPR (2014)
39. Kocasari, U., Dirik, A., Tiftikci, M., Yanardag, P.: StyleMC: multi-channel based fast text-guided image generation and manipulation. In: WACV (2022)
40. Lei, C., Liu, D., Li, W., Zha, Z.J., Li, H.: Comparative deep learning of hybrid representations for image recommendations. In: CVPR (2016)
41. Lewis, K.M., Varadharajan, S., Kemelmacher-Shlizerman, I.: TryOnGAN: body-aware try-on via layered interpolation. ACM TOG **40**(4), 1–10 (2021)
42. Li, Z., et al.: Animated 3D human avatars from a single image with GAN-based texture inference. CNG **95**, 81–91 (2021)

43. Liang, X., et al.: Deep human parsing with active template regression. PAMI **37**(12), 2402–2414 (2015)
44. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping GAN: a unified framework for human motion imitation, appearance transfer and novel view synthesis. In: ICCV (2019)
45. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: CVPR (2016)
46. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: CVPR (2019)
47. Liu, Z., Yan, S., Luo, P., Wang, X., Tang, X.: Fashion landmark detection in the wild. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 229–245. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_15
48. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: NeurIPS (2017)
49. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. In: CVPR (2018)
50. Ma, Q., et al.: Learning to dress 3D people in generative clothing. In: CVPR (2020)
51. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV (2017)
52. Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., Malossi, C.: BaGAN: data augmentation with balancing GAN. arXiv preprint arXiv:1803.09655 (2018)
53. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: ICCV (2017)
54. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C.: PULSE: self-supervised photo upsampling via latent space exploration of generative models. In: CVPR (2020)
55. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR (2018)
56. Neuberger, A., Borenstein, E., Hilleli, B., Oks, E., Alpert, S.: Image based virtual try-on network from unpaired data. In: CVPR (2020)
57. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: ICCV (2019)
58. Niemeyer, M., Geiger, A.: GIRAFFE: representing scenes as compositional generative neural feature fields. In: CVPR (2021)
59. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: StyleSDF: high-resolution 3D-consistent image and geometry generation. In: CVPR (2022)
60. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: StyleCLIP: text-driven manipulation of StyleGAN imagery. In: ICCV (2021)
61. Patel, C., Liao, Z., Pons-Moll, G.: TailorNet: predicting clothing in 3D as a function of human pose, shape and garment style. In: CVPR (2020)
62. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: CVPR (2017)
63. Pumarola, A., Sanchez-Riera, J., Choi, G., Sanfeliu, A., Moreno-Noguer, F.: 3DPeople: modeling the geometry of dressed humans. In: CVPR (2019)
64. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: ImageBERT: cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966 (2020)
65. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)

66. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2016)
67. Ramesh, A., et al.: Zero-shot text-to-image generation. In: ICML (2021)
68. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM TOG **42**(1), 1–13 (2022)
69. Sarkar, K., Golyanik, V., Liu, L., Theobalt, C.: Style and pose control for image synthesis of humans from a single monocular view. arXiv preprint arXiv:2102.11263 (2021)
70. Sarkar, K., Mehta, D., Xu, W., Golyanik, V., Theobalt, C.: Neural re-rendering of humans from a single image. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 596–613. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_35
71. Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., Varshney, K.R.: Fairness GAN: generating datasets with fairness properties using a generative adversarial network. IBM JRD **63**(4/5), 3-1 (2019)
72. Sauer, A., Schwarz, K., Geiger, A.: StyleGAN-XL: scaling StyleGAN to large diverse datasets. ACM TOG (2022)
73. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: GRAF: generative radiance fields for 3D-aware image synthesis. In: NeurIPS (2020)
74. Serengil, S.I., Ozpinar, A.: HyperExtended LightFace: a facial attribute analysis framework. In: ICEET (2021)
75. Shen, Y., Yang, C., Tang, X., Zhou, B.: InterFaceGAN: interpreting the disentangled face representation learned by GANs. PAMI **44**(4), 2004–2018 (2020)
76. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in GANs. In: CVPR (2021)
77. Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N.: Deformable GANs for pose-based human image generation. In: CVPR (2018)
78. Song, D., Tong, R., Chang, J., Yang, X., Tang, M., Zhang, J.J.: 3D body shapes estimation from dressed-human silhouettes. In: CGF (2016)
79. Tewari, A., et al.: PIE: portrait image embedding for semantic control. ACM TOG **39**(6), 1–14 (2020)
80. Tewari, A., et al.: StyleRig: rigging StyleGAN for 3D control over portrait images. In: CVPR (2020)
81. Toutouh, J., Hemberg, E., O'Reilly, U.-M.: Data dieting in GAN training. In: Iba, H., Noman, N. (eds.) Deep Neural Evolution. NCS, pp. 379–400. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-3685-4_14
82. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for StyleGAN image manipulation. ACM TOG **40**(4), 1–14 (2021)
83. Tzelepis, C., Tzimiropoulos, G., Patras, I.: WarpedGANSpace: finding non-linear RBF paths in GAN latent space. In: ICCV (2021)
84. Véges, M., Lőrincz, A.: Absolute human pose estimation with depth prediction network. In: IJCNN (2019)
85. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 607–623. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_36
86. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity GAN inversion for image attribute editing. In: CVPR (2022)
87. Wang, T., Zhang, T., Lovell, B.: Faces a la carte: text-to-face generation via attribute disentanglement. In: WACV (2021)

88. Wu, C., Li, H.: Conditional transferring features: scaling GANs to thousands of classes with 30% less high-quality data for training. In: IJCNN (2020)
89. Wu, Q., Li, L., Yu, Z.: TextGAIL: generative adversarial imitation learning for text generation. In: AAAI (2021)
90. Wu, Z., Lischinski, D., Shechtman, E.: StyleSpace analysis: disentangled controls for StyleGAN image generation. In: CVPR (2021)
91. Xia, W., Yang, Y., Xue, J.H., Wu, B.: TediGAN: text-guided diverse face image generation and manipulation. In: CVPR (2021)
92. Xu, D., Yuan, S., Zhang, L., Wu, X.: FairGAN: fairness-aware generative adversarial networks. In: IEEE BigData (2018)
93. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: generative 3D human shape and articulated pose models. In: CVPR (2020)
94. Xu, T., et al.: AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: CVPR (2018)
95. Xu, Y., et al.: TransEditor: transformer-based dual-space GAN for highly controllable facial editing. In: CVPR (2022)
96. Yildirim, G., Jetchev, N., Vollgraf, R., Bergmann, U.: Generating high-resolution fashion model images wearing custom outfits. In: ICCVW (2019)
97. Zhang, D., Khoreva, A.: Progressive augmentation of GANs. In: NeurIPS (2019)
98. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient GAN training. In: NeurIPS (2020)
99. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: ICCV (2015)