Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

# Identity-Consistent Transfer Learning of Portraits for Digital Apparel Sample Display

**Luyuan Wang*[1] | Yiqian Wu*[1] | Yong-Liang Yang[2] | Chen Liu[1] | Xiaogang Jin[1]**

[1]State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China

[2]Department of Computer Science, University of Bath, Bath, UK

**Correspondence**
Xiaogang Jin is the corresponding author.
Email: jin@cad.zju.edu.cn

**Present address**
State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China.

Funding Information
This research was supported by the Ningbo Major Special Projects of the "Science and Technology Innovation 2025" (Grant No. 2023Z143) and the Key R&D Program of Zhejiang (No. 2023C01047); Research Council UK grant CAMERA (EP/M023281/1, EP/T022523/1).

## Abstract

The rapid development of the online apparel shopping industry demands innovative solutions for high-quality digital apparel sample displays with virtual avatars. However, developing such displays is prohibitively expensive and prone to the well-known "uncanny valley" effect, where a nearly human-looking artifact arouses eeriness and repulsiveness, thus affecting the user experience. To effectively mitigate the "uncanny valley" effect and improve the overall authenticity of digital apparel sample displays, we present a novel photo-realistic portrait generation framework. Our key idea is to employ transfer learning to learn an identity-consistent mapping from the latent space of rendered portraits to that of real portraits. During the inference stage, the input portrait of an avatar can be directly transferred to a realistic portrait by changing its appearance style while maintaining the facial identity. To this end, we collect a new dataset, **D**az-**R**endered-**F**aces-**HQ** (*DRFHQ*), specifically designed for rendering-style portraits. We leverage this dataset to fine-tune the StyleGAN2-*FFHQ* generator, using our carefully crafted framework, which helps to preserve the geometric and color features relevant to facial identity. We evaluate our framework using portraits with diverse gender, age, and race variations. Qualitative and quantitative evaluations, along with ablation studies, highlight our method's advantages over state-of-the-art approaches.

**KEYWORDS**
digital apparel, portrait authenticity, uncanny valley effect, transfer learning, StyleGAN

## 1 | INTRODUCTION

Over the past decade, online clothing shopping has led to the development of digital garment examples created with professional design softwares[1,2,3]. The garment, the human body, and the face dominate the content of display images. However, creating visually convincing and photorealistic display images of these samples involves a range of laborious and expertise-intensive operations, such as high-quality garment modeling, high-fidelity human modeling, material creation, and lighting setup. The appearance of avatars' faces, in particular, has a significant impact on the authenticity of a display image and also poses a particular challenge to create the display image realistically. To generate high-fidelity faces, multi-view stereo system-based[4] and light-stage-based facial appearance capture methods[5,6] have been proposed. However, these approaches are highly expensive and time-consuming. Despite the superior quality of these faces, they often contain subtle unrealistic details that are immediately noticeable because humans are innately sensitive to such details when perceiving faces. These unrealistic flaws can suddenly shift a person's response to the avatars from empathy to eerie, frightening, or revulsion, a phenomenon known as the "uncanny valley" effect[7,8], which significantly detracts from the user experience. The rise of deep learning, in particular Generative Adversarial Networks (GANs)[9], has inspired researchers to develop high-quality face generation methods[10]. In recent years, StyleGAN[11] and its variants[12], along with their inversion techniques[13], have paved the way for the semantic manipulation[14] of photo-realistic portraits. The existing methods that can improve the realism of avatar faces[15] are all based on projecting the rendering-style faces into the pretrained StyleGAN2-*FFHQ* generator, thanks to its high generation quality and diversity. Garbin et al.[16] matches a non-photorealistic portrait to a latent code of the pretrained StyleGAN2 generator while maintaining pose, expression, hair,

* Contribute equally to this work.

  


**FIGURE 1** Given digital apparel sample display images as input (a, c), our method can effectively improve the realism of avatars' faces (b, d). The gray boxes highlight the faces. We also apply our method to rendering-style portraits (e), producing photo-realistic results (f).

and lighting consistency. Despite the attempt to adapt to the real face domain, their method requires a substantial amount of processing time for each image. Furthermore, since the input is out of the domain of the pretrained model, the output often has artifacts such as distortion and identity inconsistency. Chandran et al.[15] project high-quality yet incompletely rendered facial skin into the latent space of StyleGAN2, generating temporally-coherent and photo-realistic portraits. Nevertheless, their method is more of an inpainting process for the missing face components, such as hair, eyes, and mouth interior. Also, the output images still retain the rendering style thus lack authenticity.

The limitations of the existing works motivate us to present a novel StyleGAN-based portrait generation framework to increase the authenticity of digital apparel display. We propose a transfer-learning-based approach to establish the correlation between portrait images with different styles. The key idea is to develop an identity-consistent fine-tuning method that results in a rendering-style generator with facial identities matching those of the realistic-style StyleGAN2-*FFHQ* generator. We treat a latent code in the *W+* latent space of a portrait as an implicit representation of both portrait style and identity. While the portrait style can be either a rendering style or a realistic style corresponding to the two generators, the portrait identity is shared in-between. That is, if we project a rendered portrait into the rendering-style generator's *W+* latent space, the realistic-style StyleGAN2 generator can interpret the resulting latent code as a realistic-style portrait with the rendered portrait's facial identity. We find that by doing so, the rendering-style can be effectively removed from the final output, and the facial identity can be preserved without distortion. Based on this principle, we first collect a new dataset of rendering-style portraits, **D**az-**R**endered-**F**aces-**HQ** (*DRFHQ*). Inspired by StyleGAN2-ada[17], we use *DRFHQ* to finetune the pretrained StyleGAN2-*FFHQ* generator, resulting in a rendering-style StyleGAN2-*DRFHQ* generator. During finetuning, we constrain with sketches and color to help the new generator maintain facial identities. Then we perform latent code optimization to project the input rendering-style portrait into StyleGAN2-*DRFHQ*'s latent space. Finally, we feed the resulting latent code into the pretrained StyleGAN2-*FFHQ* generator, yielding a photo-realistic portrait with preserved facial identity. Extensive evaluations demonstrate that our work is capable of generating plausible results for digital apparel sample display.

In summary, our work makes the following contributions:

- We present the first portrait generation framework to overcome the "uncanny valley" effect in digital apparel sample displays.
- Based on a new high-quality rendering-style portrait dataset (*DRFHQ*), we propose a novel transfer-learning-based approach to correlate portraits with different styles in the learnt latent space while preserving facial identity.

# 2 | RELATED WORK

## 2.1 | Portrait Synthesis.

Human face modeling and rendering is a crucial and active research topic for applications in the entertainment, film, and television industries. Most physically-based rendering methods require a multi-view stereo system to reconstruct pore-level geometry and skin reflectance properties[4]. To capture detailed human faces, a number of light-stage-based approaches[6] have been developed based on the seminal work for facial appearance capturing and reconstruction[5]. Although the photo-realistic renderings of avatars are almost indistinguishable from real humans, the "uncanny valley" effect occurs when an anomaly is revealed from their seemingly realistic appearance[18]. Researchers have suggested methods to measure the "uncanny valley" effect[19], however, it is difficult to eliminate such an unpleasant effect in traditional rendering. Since their introduction in 2020, neural radiance fields (NeRF)[20] have spawned a slew of downstream applications, including face synthesis. However, existing face modeling and rendering methods still struggle to produce photo-realistic results that can avoid the "uncanny valley" effect. The introduction of generative adversarial networks (GANs)[9] sparks an increasing number of face synthesis models[10,12]. Among these works, StyleGAN[12] is mostly favored due to its synthesis quality and manipulation ability, and serves as an inspiration for many downstream works[21].

## 2.2 | Face Style Transfer using StyleGAN.

Portrait style transfer using StyleGAN is also related to our work. Pinkney and Adler[22] use a resolution-dependent method to interpolate different styles at appearance and geometry levels. Wu et al.[23] conduct a thorough investigation into the properties of aligned StyleGAN and use their findings to investigate potential applications such as cross-domain image morphing and zero-shot vision tasks. In addition to example images, StyleGAN-NADA[24] uses text prompt as input to stylize portraits with the help of a pretrained CLIP model. This line of research has been expanded to videos[25] to achieve consistent results in a sequence. Sang et al.[26] also attempt to create stylized and editable 3D models directly from users' avatars. However, the above methods are intended to generate stylized portraits from real photos, whereas our work aims at the opposite: transfer the "rendering-style" of the rendered portraits into the "realistic-style" of the results that are indistinguishable from real portraits.

## 2.3 | Face Realism Improvement using StyleGAN.

Improving the realism of rendered faces is still a challenging issue. Garbin et al.[16] propose a zero-shot image projection algorithm that requires no training data to find the latent code that most closely matches the input face. Their objective is the most similar to ours. Chandran at al.[15] use a multi-frame consistent method to project the traditional incomplete face rendering results into latent space to achieve realistic rendering and animation of a full-head portrait. Despite generating realistic full-head portraits, their primary goal is to inpaint the missing components. As a result, their method preserves the input rendered skin but is incapable of improving the faces' authenticity. The StyleGAN encoders[27] and some optimization-based methods[28] can project the rendered faces into StyleGAN's latent space. However, the rendered faces are far outside the domain of the real faces, thus resulting in distortion and artifacts or maintaining the "rendering-style". Different from these methods, we focus on producing realistic portraits for digital apparel sample display while preserving the facial identity.

# 3 | METHOD

Our goal is to improve the authenticity of the digital apparel sample display by replacing the avatar's portrait with a realistic one while maintaining the avatar's facial identity. At the same time, we leave all other portions of the body unchanged to retain the appearance of the garment.

Fig. 2 demonstrates the key idea of our approach. We conduct portrait replacement in the latent space by employing latent code that implicitly represents portrait style and identity as the interface in-between. As shown in Fig. 2 (a), we establish identity-consistent transfer learning on the StyleGAN generator of realistic portraits ($G_{real}$), resulting in a fine-tuned generator ($G_{rendering}$) of portraits with a different style, i.e., the "rendering" style. The transfer learning is performed in a way that given a single latent code in the $W+$ latent space, the portrait identity can be well preserved in both generators, only the portrait style is interpreted differently as "realistic-style" by $G_{real}$ and "rendering-style" by $G_{rendering}$. In other words, the same latent code can

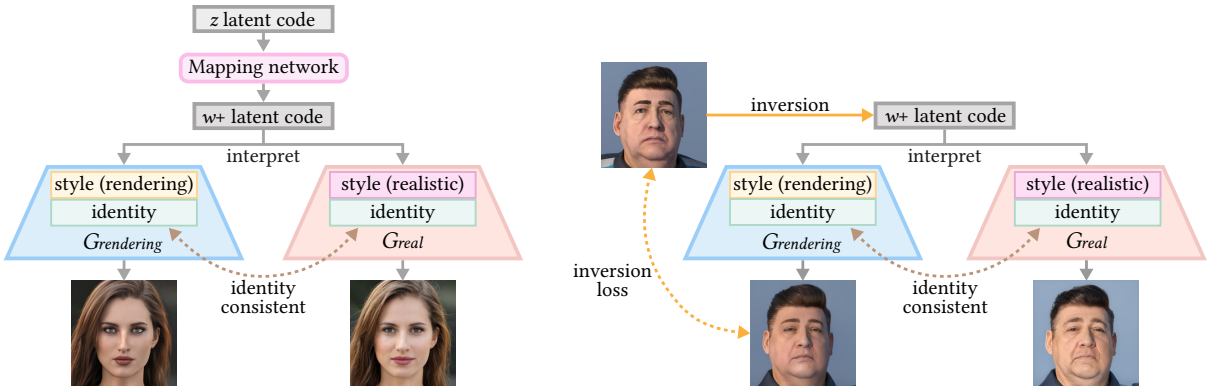


**FIGURE 2** The central idea of our method.

generate two portraits with distinct styles but matched identity. Unlike the style change in the fine-tuning process, during the inference phase (see Fig. 2 (b)), we aim to "invert" the style of the input portrait from "rendering" to "realistic". We begin by applying GAN inversion to obtain the avatar's latent code in the $W+$ latent space of $G_{rendering}$. The latent code is then fed into $G_{real}$ to adapt to the realistic style while preserving identity. In the end, we achieve the final result - a photorealistic portrait with the identity of the input avatar.

The rest of the section is organized as follows. We begin by introducing *DRFHQ*, a new high-quality rendering-style portrait dataset used for transfer learning (Sec. 3.1). Then we elaborate our transfer learning strategy, which initializes $G_{rendering}$ with the weights of $G_{real}$ and fine-tunes $G_{rendering}$ with a different style while minimizing other irrelevant changes (Sec. 3.2). Finally, we present how we increase the authenticity of digital apparel sample display in the inference phase (Sec. 3.3).

## 3.1 | Daz-Rendered-Faces-HQ dataset

We create **D**az-**R**endered-**F**aces-**HQ** (*DRFHQ*), a dataset that comprises high-quality rendering-style portrait images, by collecting daz3d.com's gallery[29]. *DRFHQ* contains 11,399 high-quality PNG images in $1024^2$ resolution, with a wide range of gender, age, pose, race, hairstyle, etc. We first align and crop the raw images using Dlib[30] according to the preprocessing method of *FFHQ*, then manually filter the aligned images. Due to copyright restrictions, we cannot release the collected images but will provide the corresponding URLs as an alternative. Although several publicly available rendering-style datasets exist[31,32,33,34], their face resolution is insufficient for high-quality digital apparel sample display[31,34], or they only contain a small number of rendered faces[32,33], or they are rendered using a small number of face models (100 different identities)[34]. *DRFHQ* is the first high-quality rendering-style dataset with a face region resolution of $1024^2$ that can be extended to downstream tasks, to the best of our knowledge.

## 3.2 | Identity-Consistent Transfer Learning

Inspired by StyleGAN-ada[17], we use *DRFHQ* to fine-tune the generator $G_{rendering}$ initialized with the weights of the pretrained StyleGAN2-*FFHQ* generator $G_{real}$, resulting in a new stylized generator StyleGAN2-*DRFHQ* capable of producing rendering-style portraits. To replace the rendered face in the digital apparel sample display image with a realistic face while meeting the designer's preference, we want to keep the facial identity unchanged to avoid any unnatural artifacts. However, simply fine-tuning $G_{rendering}$ leads to large facial identity deviations in the fine-tuned latent space compared to the original. To address this issue, we use two additional losses during the fine-tuning process to constrain the facial identity. The training pipeline is illustrated in Fig. 3.

Our idea is to use the same latent code in $W+$ latent space to implicitly represent the rendered face and its realistic face replacement, hence $G_{rendering}$ and $G_{real}$ are required to share the same $W+$ latent space. To do this, we freeze the mapping network during fine-tuning, resulting in a single latent code $z$ in $Z$ latent space being mapped to the same latent code $w+ \in W+$ of $G_{rendering}$ and $G_{real}$. We will omit the unmodified mapping network in the remainder of this section and use $w+$ as the latent code.

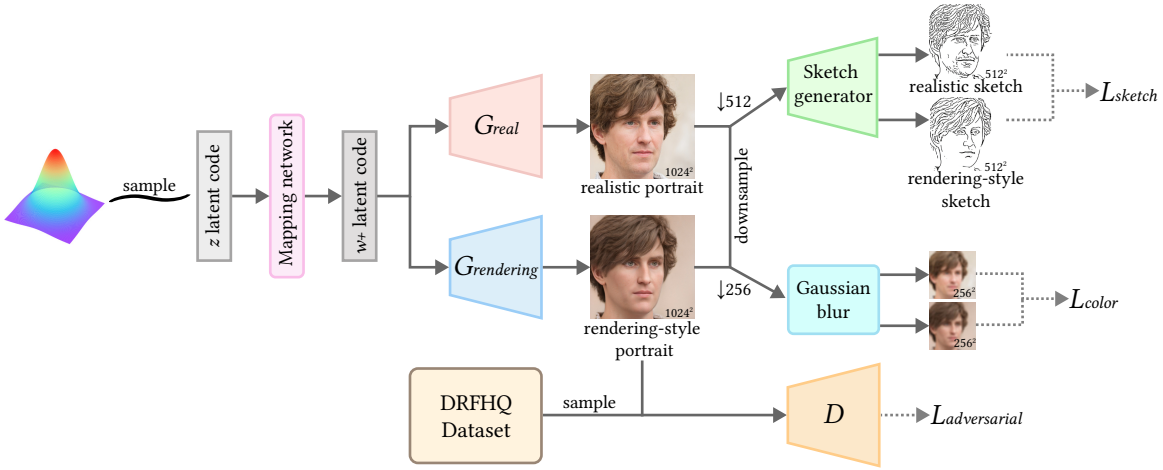


**FIGURE 3** An overview of our identity-consistent transfer learning network's training stage.

**Sketch loss.** Inspired by DeepFaceEditing[35], the geometric features of the face can be well represented by sketches. Therefore, we add the following L1 loss function:

$$L_{sketch} = \|S(G_{real}(w+) \downarrow_{512}) - S(G_{rendering}(w+) \downarrow_{512})\|_1, \tag{1}$$

where $G_{rendering}$ is to be fine-tuned and initialized by the pretrained $G_{real}$, $S$ is the pretrained sketch extractor in DeepFaceEditing[35] model, and $\downarrow_{512}$ denotes the interpolation operation that downsamples the images to $512 \times 512$. According to Eq. 1, the output of $G_{real}$ and $G_{rendering}$ are fed into $S$ separately to obtain two face sketches, and the geometric contours of the two faces are constrained to be as similar as possible by using the L1 norm.

**Color loss.** To preserve the portrait color during transfer learning, we propose a color loss at the perceptual level based on the LPIPS loss[36]. However, LPIPS captures the facial appearance similarity, including texture and style-related details, preventing the generator from learning rendering-style. Inspired by[16], we solve this problem by removing the appearance details from the images. Specifically, we first downsample the images to $256 \times 256$ and apply Gaussian blur, then feed the images into the VGG16 network to compute the LPIPS loss:

$$L_{color} = LPIPS(B(G_{real}(w+) \downarrow_{256}), B(G_{rendering}(w+) \downarrow_{256}), \tag{2}$$

where $B$ is the Gaussian blur operation with $kernel = 13$ and $\sigma = 10$, and $\downarrow_{256}$ denotes the interpolation operation that downsamples the images to $256 \times 256$. Our objective loss function used in fine-tuning is the weighted sum of the following losses:

$$L_G = L_{origin} + \lambda_s L_{sketch} + \lambda_c L_{color}, \tag{3}$$

where we empirically set $\lambda_s = 5 \times 10^{-6}$ and $\lambda_c = 3.75 \times 10^3$. $L_{origin}$ is the original loss of StyleGAN-ada.

## 3.3 | Inference

In the inference phase, we use the latent optimization[37] inversion method to project the rendered portrait $x$ onto $G_{rendering}$'s latent space. As we aim for the least distortion, we optimize in the $W+$ latent space, which has greater expressive potential:

$$w+^*, n^* = \arg\min_{w+,n} \lambda_n L_n(n) + LPIPS(x, G_{rendering}(w+, n)), \tag{4}$$

where $G_{rendering}(w+, n)$ is image generated by $G_{rendering}$ with noise $n$, $L_n$ is a noise regularization term, and $\lambda_n = 1e5$. We initialize $w+$ as the average latent code in the $W+$ latent space and use a 500-step optimization to get $w+^*$. Finally, we input the resulting latent code $w+^*$ to $G_{real}$, yielding a photo-realistic portrait. We do not employ the optimized noise $n^*$ here because the regularization term $L_n$ prevents the noise vector from influencing the final result.

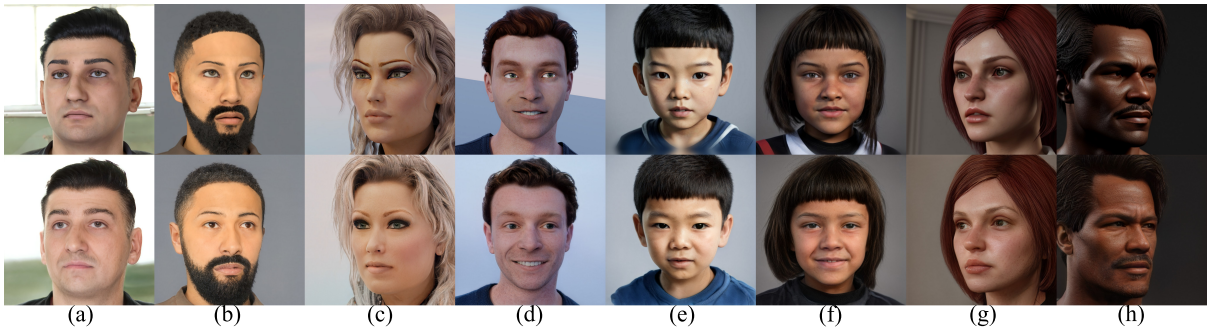


**FIGURE 4** Diverse photo-realistic results (bottom) generated by our method.

## 4 | EXPERIMENTS

This section showcases the outcomes of our identity-consistent transfer learning framework. We present the results of our approach as applied to a series of rendering-style portraits. In Figs. 1(e)(f) and 4, we display a variety of results that span various genders, ages, and races, effectively illustrating how our approach can adapt across diverse data sources (e.g. Diverse Human Faces dataset[34], internet images[38], and rendering-style images generated using Stable Diffusion[39]. Additionally, we also showcase some examples where we stitch the generated realistic faces back onto the original garment display renderings (Fig. 1(a)(b), and Fig. 2 in the supplementary material). Our generated realistic faces can easily blend in with the rendered garment and virtual avatar bodies with only minor post-processing (see the supplementary material). The adoption of our method can significantly enhance the overall authenticity of apparel display renderings. In sum, our method effectively overcomes the "uncanny valley" effect (see Sec. 4.2.1) by largely improving the authenticity of rendered faces while avoiding portrait infringement liability due to using generated faces. Furthermore, it preserves the facial identity, aligning with the designer's preference.

### 4.1 | Implementation Details

**Networks.** We use the StyleGAN2-ada architecture[17] as the backbone for our rendering-style generator. StyleGAN2-*FFHQ* is the official pretrained model of StyleGAN2-ada on the *FFHQ* dataset. We use the training parameters (batch size of 32, learning rate of $2e-3$) provided in the stylegan2 config of StyleGAN2-ada to finetune StyleGAN2-*FFHQ* while freezing the weights of the ToRGB layers and the mapping network. We only update $G_{rendering}$ and the discriminator, while $G_{real}$ and the sketch extractor are fixed. The training dataset is amplified with x-flips, and the fine-tuning time is about 40 minutes on 4 Tesla V100 GPUs, we stop the fine-tuning when the discriminator had seen a total of 40k real images. PyTorch is utilized to train the networks and all comparisons are conducted on a desktop PC with Intel Core i7-12700F 2.10 GHz CPU, 32GB RAM and GeForce RTX 3080Ti GPU (12GB memory). All images used in the training and testing stages have a resolution of $1024^2$. Regarding runtime performance, the average time for projecting a rendered portrait into a latent code is 27.6 seconds, with the generation of the final result only taking 0.05 seconds. All the other steps within our approach require negligible time.

**Dataset.** The fine-tuned rendering-style generator is trained using the *DRFHQ* dataset's 11,399 rendering-style portraits. The testing images in the paper are from *Diverse Human Faces*[34] dataset (Figs. 2, Fig 4 (a)(b), 11, 13), rendering-style images generated using Stable Diffusion[39] (Figs. 4 (e)-(h), 5, 7, 10), Flickr[38] (Fig. 4 (c)(d)), and CONNECT store[2] (Fig. 1) with courtesy of the authors. Specifically, we employ the fine-tuned and LoRA models based on Stable Diffusion 1.5 for generating rendering-style images.

### 4.2 | Comparison with State-of-the-Art Methods

We begin by presenting comparisons between our proposed method and state-of-the-art (SOTA) facial realism-improving methods. In Sec. 2, we mentioned that previous works[15,16] can enhance the realism of rendered faces. However, their datasets

and codes are not publicly accessible. Therefore, we rely on comparisons with StyleGAN inversion methods (Sec. 4.2.1) and SDEdit (Sec. 4.2.2). Subsequently, we provide comparisons between our identity-consistent style-transfer method and SOTA style-transfer methods (Sec. 4.2.3).

### 4.2.1 | Comparison with StyleGAN inversion methods

We perform qualitative and quantitative experiments to compare our method with StyleGAN inversion methods, which project unrealistic images onto the manifold of natural images through image inversion.

**Qualitative evaluation.** To accomplish qualitative comparison, we directly project the input rendered portrait into the $W+$ latent space of StyleGAN2-*FFHQ* via StyleGAN inversion, and then compare the inversion results with our own outcomes. As illustrated in Fig. 5, we use e4e[40], pSp[41], HyperStyle[42], ReStyle[43], and latent code optimization[44], for comparison. Those encoders are trained on both *FFHQ* dataset and StyleGAN2-*FFHQ*. For ReStyle, we run testing on both e4e and pSp encoders, using the ReStyle scheme. For latent code optimization, we use the same inversion method described in Sec. 3.3 to project the input images into the $W+$ latent space of StyleGAN2-*FFHQ*. It is clear that those encoders lose many skin characteristics and produce faces with only smooth skin, which lacks realism. Furthermore, they retain the rendering style of the input images that looks unrealistic. Our method, on the other hand, produces more photo-realistic results with more natural facial details and completely changes the input image's unrealistic rendering-style appearance while maintaining facial identity consistency.


**FIGURE 5** Qualitative comparisons with state-of-the-art StyleGAN inversion methods.

**Quantitative evaluation.** To the best of our knowledge, currently there is no viable quantitative metric for assessing the authenticity of synthetic portraits. Furthermore, determining the authenticity of a portrait is largely dependent on human cognitive abilities. In light of this, we devised a user study as a quantitative experiment, with the goal of comparing the authenticity of the results produced by our proposed method to those produced by SOTA StyleGAN inversion methods. We collected ten rendered portraits and subjected them to the six StyleGAN inversion methods mentioned above (see qualitative experiments in Sec. 4.2.1) and our proposed method, respectively. We presented these ten sets of test cases sequentially to 20 participants, randomly displaying the results for authenticity comparison. Fig. 6 shows that the vast majority of our results are more realistic. This demonstrates our approach's superiority over other StyleGAN inversion methods in improving the authenticity of rendered portraits.

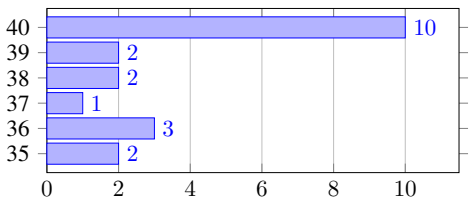


**FIGURE 6** The distribution of the user study on the authenticity comparison of the methods for improving facial realism. The $y$–axis shows the number of output portraits from our method chosen by participants (out of 10 sets), and the $x$–axis shows the number of participants. The results show that our method outperforms other methods for improving facial authenticity.

### 4.2.2 | Comparison with SDEdit

We conduct a comparison with SOTA diffusion-based method, SDEdit[45]. SDEdit projects an unrealistic image onto the manifold of natural images by adding noise and then denoising.

Given a single input, our method generates a singular result, whereas SDEdit produces stochastic results based on the random noise that is added. Consequently, conducting a fair user study as an alternative to quantitative testing poses challenges. Therefore, we opt for qualitative experiments exclusively. For each input rendering-style image, we set the hyperparameter $t_0 = 0.3$ for SDEdit and generate three randomly sampled results utilizing the pretrained latent diffusion model[39] trained on the *FFHQ* dataset at the resolution of $256^2$. As shown in Fig. 7, the results produced by SDEdit do not ensure the complete removal of the rendering-style from the input, and they also do not guarantee facial identity preserving. In contrast, our approach stably generates more photo-realistic results, showcasing enhanced natural facial details. Moreover, our approach effectively removes the unrealistic rendering-style appearance of the input image while preserving the consistency of facial identity.


**FIGURE 7** Qualitative comparison with SDEdit.

### 4.2.3 | Style Transfer

In this section, we conduct qualitative and quantitative experiments to show the effectiveness of our identity-consistent style transfer algorithm. We will show that our style-transfer approach surpasses other style transfer methods in both style transfer and facial identity preservation.

We compare our identity-consistent transfer method to the SOTA StyleGAN-based style transfer methods. Since one-shot domain adaptation methods[46,47] stylize the whole latent space using a single reference image, we cannot apply them to process our diverse testing images. Thus we make comparisons with StyleGAN-NADA[24] and AgileGAN[48].

**Qualitative evaluation.** We present a comparison between the style transfer results of our identity-consistent style transfer method and those of StyleGAN-NADA[24] and AgileGAN[48] in Fig. 8. For StyleGAN-NADA, we choose "Photo" as the source text and "Rendered avatar" as the target text. For AgileGAN, we use our *DRFHQ* dataset as the training dataset to train AgileGAN. We compare the images by column generated by different generators using the same latent code. Results show that the StyleGAN-NADA semantic guidelines are too vague to produce acceptable results. AgileGAN generates artifacts and unnatural skin color. In contrast, our approach produces rendering-style results while preserving face identity.

**Quantitative evaluation.** To evaluate the performance in transferring style to that of *DRFHQ* dataset, we utilize Fréchet Inception Distance (FID)[49] to measure the overall similarity between the distribution of synthesized images and that of the *DRFHQ* dataset (see Table 1 column 2). Besides, to evaluate the geometry and color preservation quality, we compute the FID of the synthesized rendering-style images with respect to the realistic-style *FFHQ* dataset (see Table 1 column 3). In Table 1, StyleGAN2-*DRFHQ* represents our identity-consistent model fine-tuned on our *DRFHQ* dataset, AgileGAN-*DRFHQ* represents AgileGAN[48] fine-tuned on our *DRFHQ* dataset. Since StyleGAN-NADA[24] is text-guided and not trained on our *DRFHQ* dataset, FID is for reference only. All FID scores are computed from randomly generated 50k images. Lower scores are better. Our model achieves the lowest FID as shown in Table 1, indicating that our StyleGAN2-*DRFHQ* model is better at both style transfer and facial identity preservation.

To further assess the performance in facial identity preservation, we utilize a pretrained CurricularFace network[50] to compute identity similarity during facial style transfer. Specifically, we apply our StyleGAN2-*DRFHQ* model, AgileGAN-*DRFHQ* model, and StyleGAN-NADA model to convert the style of 2k images from realistic to rendering respectively, we then use the


**FIGURE 8** Qualitative comparisons with SOTA style transfer methods.

**TABLE 1** FID and Identity similarity score for different StyleGAN-based style transfer methods and datasets.

| Algorithm | *DRFHQ*↓ | *FFHQ*↓ | Identity Similarity ↑ |
|---|---|---|---|
| StyleGAN2-*DRFHQ* (Ours) | **24.5** | **16.3** | **0.57** |
| AgileGAN-*DRFHQ* | 62.5 | 83.5 | 0.14 |
| StyleGAN-NADA | 49.9 | 53.8 | 0.34 |

CurricularFace network to measure facial identity. As shown in Table 1 column 4, our StyleGAN2-*DRFHQ* model exhibits superior performance in preserving facial identity during the process of style transfer. Higher scores are better.

## 4.3 | Ablation Studies

We perform ablation studies to validate the effectiveness of different components of our work. We first evaluate the two proposed losses (Sec. 4.3.1), then our transfer-learning-based framework (Sec. 4.3.2), the employed inversion method (Sec. 4.3.3), and finally our new high-quality rendering-style portrait dataset (Sec. 4.3.4).

### 4.3.1 | Losses

We define StyleGAN2-*FFHQ* generator fine-tuned on our *DRFHQ* dataset without sketch and color constraints as the baseline. As shown in Fig. 9, we feed the same latent codes into generator variants and compare the results.


**FIGURE 9** Exemplars of the baseline and ours.

**Sketch loss.** Without the sketch constraint, the identity of the face generated by the baseline differs significantly from that of StyleGAN2-*FFHQ*, thus largely affecting facial identity consistency. Thanks to $L_{color}$, the generator trained without $L_{sketch}$ generates portraits that better maintain the identity. However, the semantic information cannot be well preserved due to the

downsampling and blurring of the images fed into the VGG16 network (see the details of the facial expressions and wrinkles in the images). In contrast, $L_{sketch}$ helps to keep detailed facial structure and semantics in our full model.

**Color loss.** Compared to generators trained without color constraint, those trained with color constraint can better preserve the color and lighting of the portraits generated by StyleGAN2-*FFHQ*.

### 4.3.2 | Framework

Although our sketch loss and color loss provide strong guidance for identity preservation, the proposed losses alone are not enough to generate satisfactory results without our carefully designed framework. Note that our framework includes model fine-tuning with our proposed losses, followed by inversion and generation to produce the final results. As a baseline, we directly project input rendered images into the realistic portrait latent space (StyleGAN2-*FFHQ*) using our proposed losses as guidance for latent code optimization.

As shown in Fig. 10, we compare the baseline result to ours. It can be seen that the baseline produces overly smooth results, while our framework generates more realistic result. Actually, the sketches and downsampled blurry images in the proposed losses can provide key identity information but at a coarse level, thus leading to smooth results that lack details. In contrast, our framework uses a ∼10k dataset to fine-tune the StyleGAN2-*FFHQ* model, which is pretrained on a ∼70k dataset. Both large-scale datasets are rich in face features at different levels. The fine-tuning process can effectively model the delicate details of the rendering-style faces in $G_{render}$, allowing to achieve more realistic results when transferring to $G_{real}$.


**FIGURE 10** Exemplars of the ablation study of the baseline method and ours.

### 4.3.3 | Inversion

In our framework, we use the latent code optimization described by Roich et al.[44] as our inversion method during inference. We compare it to the following cutting-edge inversion approaches: e4e[40], ReStyle scheme on e4e (ReStyle-e4e)[43], and II2S[51].

For e4e and ReStyle-e4e, we fine-tune their encoders pretrained on the *FFHQ* dataset using our *DRFHQ* dataset. Then, we input the rendered images into these fine-tuned encoders, respectively. For II2S, we use it to directly project input rendered images into $G_{rendering}$'s latent space. Finally, we feed these latent codes into $G_{real}$ to yield the final results for comparison. As shown in Fig. 11, e4e changes facial identity and gender (the first row). ReStyle-e4e lacks facial details, and II2S modifies input image attributes (glasses appear in the second row of II2S). In contrast, our inversion method surpasses all others.


**FIGURE 11** Exemplars of the ablation study of different inversion methods.

### 4.3.4 | Dataset

To validate the efficacy of our high-quality rendering-style portrait dataset, *DRFHQ*, in enhancing facial realism, we qualitatively and quantitatively compare it with the *Diverse Human Faces* dataset[34]. To this end, we replace our *DRFHQ* dataset with *Diverse Human Faces* dataset during generator fine-tuning, while maintaining method consistency.


**FIGURE 12** Exemplars of the ablation study of the *Diverse Human Faces* dataset and our *DRFHQ* dataset.

**Qualitative evaluation.** We enhance facial realism in rendered images using two frameworks: one based on the *Diverse Human Faces* dataset and the other on our *DRFHQ* datasets. Note that the input rendered portraits for inference are not part of either dataset. As shown in Fig. 12, our *DRFHQ* dataset-based framework achieves photorealism and facial identity consistency, while the *Diverse Human Faces* dataset-based framework exhibits greater disparities in geometry, color and realism. The input rendered images are generated by Stable Diffusion[39] model.

We attribute this phenomenon to the limited diversity of the *Diverse Human Faces* dataset, which consists of ∼7k images (after aligning and cropping) but only portrays 100 distinct identities. In contrast, our high-quality *DRFHQ* dataset contains ∼10k high-quality images with diverse attributes like identity, gender, age, pose, race, hairstyle, lighting, etc. This diversity effectively models the delicate rendering-style facial details during fine-tuning, leading to more realistic inference outcomes.

**Quantitative evaluation.** For quantitative evaluation, we employ LPIPS loss[36] and L2 loss to assess dataset performance in information preservation.We compute the two losses from 150 pairs of images for the *Diverse Human Faces* dataset-based and our *DRFHQ* dataset-based frameworks. Lower values indicate better performance. Table 2 demonstrates that our *DRFHQ* dataset outperforms the *Diverse Human Faces* dataset in both metrics, indicating superior overall information preservation.

**TABLE 2** Mean LPIPS and L2 losses.

| Dataset | LPIPS↓ | L2↓ |
|---|---|---|
| *DRFHQ* (Ours) | **0.135** | **0.048** |
| *Diverse Human Faces* | 0.176 | 0.062 |

## 5 | LIMITATIONS AND FUTURE WORK

Our method has some limitations. When the input faces contain accessories such as unique glasses and hats, our model's results have visible inconsistencies with the original images (Fig. 13(a)). This is due to the lack of corresponding relevant semantics in the *FFHQ* latent space. This limitation can be addressed by enriching the diversity of photo-realistic face datasets. Our method meets the challenges to reconstruct the image backgrounds (Fig. 13(b)). We attribute this to StyleGAN's weak expressive capacity for complicated backgrounds. It can be solved by removing the generated background using the alpha matte. We notice that our approach cannot process those faces with extreme poses (Fig. 13(c)). This is caused by the imbalanced pose distribution in the training dataset (both *FFHQ* and *DRFHQ*). This can be improved by increasing the pose diversity of the dataset and retraining the StyleGAN model. Although our method can preserve the identity of the input rendered avatar, small chromatic aberration and misalignment still exist when we paste the resulting portrait back onto the full-body apparel sample display image (Fig. 13(d)). To achieve seamless integration, a lightweight post-processing (detailed in the supplementary material) of the resulting portrait is further applied.


**FIGURE 13** Example of failure cases.

## 6 | CONCLUSIONS

We present a novel identity-consistent transfer learning method that can remove the rendering-style appearance in the input portraits and generate photo-realistic portraits. Besides, we create a high-quality rendering-style portrait dataset, Daz-Rendered-Faces-HQ (*DRFHQ*), which includes 11,399 images with gender, age, pose, and race variations. To maintain the facial identity, we employ sketch and color constraints in the finetuning process of the StyleGAN2 generator on the *DRFHQ* dataset. During inference, we first leverage latent code optimization to the input rendering-style portrait, then feed the projected inversion latent code into the real-style StyleGAN2-FFHQ generator, and finally obtain the photo-realistic result with consistent identity. We have extensively validated our approach through both qualitative and quantitative experiments. In addition to digital apparel sample display, our method can be applied to various downstream tasks, including bringing the characters to life in artworks of numerous forms such as animation, sculpture, and painting. Moreover, our rendering-style *DRFHQ* dataset has the potential to motivate other creative applications such as virtual avatar synthesis and editing.

**How to cite this article:** Luyuan W, Yiqian W, Yong-Liang Y, Chen L and Xiaogang J. Identity-Consistent Transfer Learning of Portraits for Digital Apparel Sample Display. *J Comput Phys.* 2021;00(00):1–18.

## APPENDIX

## A APPLICATION IN DIGITAL SAMPLE DISPLAY

Fig. A1 shows more exemplars where we replace the original rendering style faces with our generated realistic faces in digital apparel sample display images. Input images are courtesy of Yayat Punching at the CONNECT store[2], except for the first one in row 1.

## B LIGHTWEIGHT POST-PROCESSING

As shown in Fig. B2, directly pasting the resulting portrait back onto the original rendered digital apparel display image may lead to small chromatic aberration and misalignment. To address this issue, we propose a lightweight post-processing method.

Specifically, we apply face parsing[52] to the processed resulting portrait $x_{res}$, getting the segmentation masks of skin, brows, eyes, eyeglasses, ears, nose, mouth, lips, and hair. Then we combine them as a single mask $m$. After that, we paste $x_{res}$ back onto the original rendered image $x$, getting $x'_{res}$, and paste $m$ to an empty image with the same shape as $x$, getting $m'$.

To achieve smooth results, we apply erosion and Gaussian blur to $m'$, the resulting mask with smooth boundary is denoted as $\hat{m}$. Finally, we compose the original rendered image $x$ and the intermediate image $x'_{res}$ as:

$$x_{final} = \hat{m} \odot x'_{res} + (1 - \hat{m}) \odot x, \tag{B1}$$

where $\odot$ denotes the element-wise multiplication.

**FIGURE A1** More exemplars of our method's application in digital sample display. We replace original rendered 3D avatars' faces with photo-realistic faces generated by our method. The results show that the generated photo-real faces blend in with the rendered garments and virtual avatar bodies, effectively increasing the authenticity of the digital apparel sample display images.

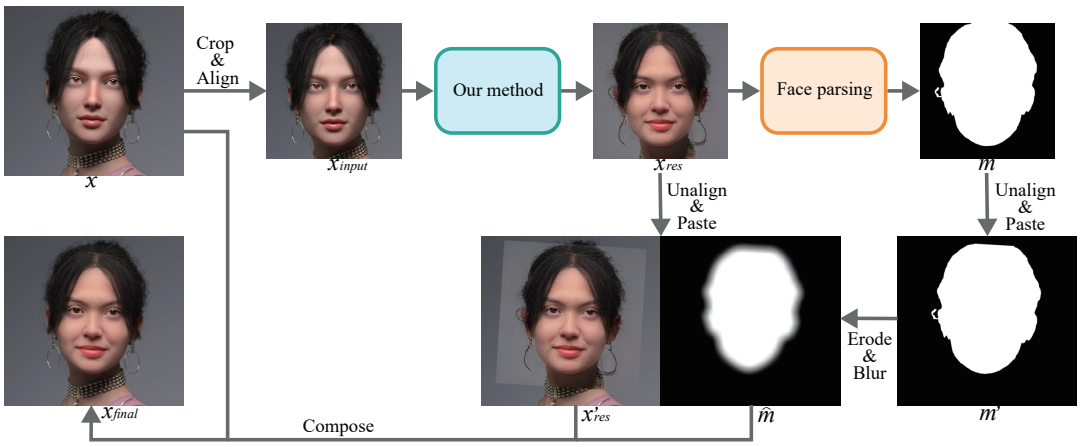


**FIGURE B2** An overview of our lightweight post-processing method.

## REFERENCES


1. Browzwear . Browzwear Solutions Pte Ltd.. Website; 2000-2024. browzwear.com.
2. CLO . CLO Virtual Fashion LLC.. Website; 2024. clo3d.com.
3. Optitex . OPTITEX. Website; 1988-2022. optitex.com.
4. Riviere J, Gotardo P, Bradley D, Ghosh A, Beeler T. Single-Shot High-Quality Facial Geometry and Skin Appearance Capture. *ACM Trans. Graph.*. 2020;39(4):81:1-81:12.
5. Debevec P, Hawkins T, Tchou C, Duiker HP, Sarokin W, Sagar M. Acquiring the Reflectance Field of a Human Face. In: SIGGRAPH '00. 2000:145–156.

6. Sun T, Xu Z, Zhang X, et al. Light Stage Super-Resolution: Continuous High-Frequency Relighting. *ACM Trans. Graph.*. 2020;39(6):260:1-260:12.
7. Mori M. Bukimi no tani [the uncanny valley]. *Energy.* 1970;7:33–35.
8. Moore RK. A Bayesian explanation of the 'Uncanny Valley'effect and related psychological phenomena. *Scientific reports.* 2012;2(1):1–5.
9. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets. In: . 27. 2014.
10. Choi Y, Uh Y, Yoo J, Ha JW. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In: 2020.
11. Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. In: 2019:4401–4410.
12. Karras T, Aittala M, Laine S, et al. Alias-Free Generative Adversarial Networks. In: . 34. 2021:852–863.
13. Liu H, Song Y, Chen Q. Delving StyleGAN Inversion for Image Editing: A Foundation Latent Space Viewpoint. In: 2023:10072-10082.
14. Pan X, Tewari A, Leimkühler T, Liu L, Meka A, Theobalt C. Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. In: SIGGRAPH '23. 2023.
15. Chandran P, Winberg S, Zoss G, et al. Rendering with style: combining traditional and neural approaches for high-quality face rendering. *ACM Trans. Graph.*. 2021;40(6):223:1–223:14.
16. Garbin SJ, Kowalski M, Johnson M, Shotton J. High Resolution Zero-Shot Domain Adaptation of Synthetically Rendered Face Images. In: 2020:220–236.
17. Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T. Training Generative Adversarial Networks with Limited Data. In: . 33. 2020:12104–12114.
18. Seyama J, Nagayama RS. The Uncanny Valley: Effect of Realism on the Impression of Artificial Human Faces. *Presence Teleoperators Virtual Environ.*. 2007;16(4):337–351.
19. Ho C, MacDorman KF. Measuring the Uncanny Valley Effect - Refinements to Indices for Perceived Humanness, Attractiveness, and Eeriness. *Int. J. Soc. Robotics.* 2017;9(1):129–139.
20. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In: . 12346 of *Lecture Notes in Computer Science*. 2020:405–421.
21. Xiong Z, Kang D, Jin D, et al. Get3DHuman: Lifting StyleGAN-Human into a 3D Generative Model Using Pixel-Aligned Reconstruction Priors. In: 2023:9287-9297.
22. Pinkney JNM, Adler D. Resolution Dependent GAN Interpolation for Controllable Image Synthesis Between Domains. *CoRR.* 2020;abs/2010.05334.
23. Wu Z, Nitzan Y, Shechtman E, Lischinski D. StyleAlign: Analysis and Applications of Aligned StyleGAN Models. *arXiv preprint arXiv:2110.11323.* 2021.
24. Gal R, Patashnik O, Maron H, Bermano AH, Chechik G, Cohen-Or D. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Trans. Graph.*. 2022;41(4):141:1–141:13.
25. Liu P, Song L, Zhang D, et al. Emo-Avatar: Efficient Monocular Video Style Avatar through Texture Rendering. 2024.
26. Sang S, Zhi T, Song G, et al. AgileAvatar: Stylized 3D Avatar Creation via Cascaded Domain Bridging. In: 2022:23:1–23:8.
27. Pehlivan H, Dalva Y, Dundar A. StyleRes: Transforming the Residuals for Real Image Editing With StyleGAN. In: 2023:1828-1837.
28. Abdal R, Qin Y, Wonka P. Image2StyleGAN++: How to Edit the Embedded Images?. In: 2020:8293–8302.
29. Productions D. Daz Productions Inc.. Website; 2024. www.daz3d.com/gallery.
30. Kazemi V, Sullivan J. One Millisecond Face Alignment with an Ensemble of Regression Trees. In: 2014:1867–1874.
31. Wood E, Baltrušaitis T, Hewitt C, Dziadzio S, Cashman TJ, Shotton J. Fake It Till You Make It: Face Analysis in the Wild Using Synthetic Data Alone. In: 2021:3681-3691.
32. Liu M, Li Q, Qin Z, Zhang G, Wan P, Zheng W. BlendGAN: Implicitly GAN Blending for Arbitrary Stylized Face Generation. In: . 34. 2021:29710–29722.
33. Oliver MM, Amengual Alcover E. UIBVFED: Virtual facial expression dataset. *Plos one.* 2020;15(4):e0231266.
34. AI S. Synthesis AI. Website; 2022. https://opensynthetics.com/dataset/diverse-human-faces-dataset/.
35. Chen SY, Liu FL, Lai YK, et al. DeepFaceEditing: Deep Face Generation and Editing with Disentangled Geometry and Appearance Control. *ACM Trans. Graph.*. 2021;40(4):90:1-15.
36. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: 2018:586–595.
37. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and Improving the Image Quality of StyleGAN. In: 2020:8107–8116.
38. Flickr . Flickr. Website; 2024. flickr.com.
39. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis With Latent Diffusion Models. In: 2022:10684-10695.
40. Tov O, Alaluf Y, Nitzan Y, Patashnik O, Cohen-Or D. Designing an encoder for StyleGAN image manipulation. *ACM Trans. Graph.*. 2021;40(4):133:1–133:14.
41. Richardson E, Alaluf Y, Patashnik O, et al. Encoding in Style: A StyleGAN Encoder for Image-to-Image Translation. In: 2021:2287–2296.
42. Alaluf Y, Tov O, Mokady R, Gal R, Bermano A. HyperStyle: StyleGAN Inversion with HyperNetworks for Real Image Editing. In: 2022:18490–18500.
43. Alaluf Y, Patashnik O, Cohen-Or D. ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement. In: 2021:6691–6700.
44. Roich D, Mokady R, Bermano AH, Cohen-Or D. Pivotal Tuning for Latent-Based Editing of Real Images. *ACM Trans. Graph.*. 2022;42(1):6:1-6:13.
45. Meng C, He Y, Song Y, et al. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. 2022.
46. Zhu P, Abdal R, Femiani J, Wonka P. Mind the Gap: Domain Gap Control for Single Shot Domain Adaptation for Generative Adversarial Networks. In: 2022:1–12.
47. Zhang Z, Liu Y, Han C, Guo T, Yao T, Mei T. Generalized One-shot Domain Adaptation of Generative Adversarial Networks. In: 2022.
48. Song G, Luo L, Liu J, et al. AgileGAN: stylizing portraits by inversion-consistent transfer learning. *ACM Trans. Graph.*. 2021;40(4):117:1–117:13.
49. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems*. 30. 2017.
50. Huang Y, Wang Y, Tai Y, et al. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. 2020.
51. Abdal R, Qin Y, Wonka P. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. In: 2019:4431–4440.
52. zllrunning . face-parsing.PyTorch. https://github.com/zllrunning/face-parsing.PyTorch; 2019.