Enhancing the Authenticity of Rendered Portraits with Identity-Consistent Transfer Learning

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

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

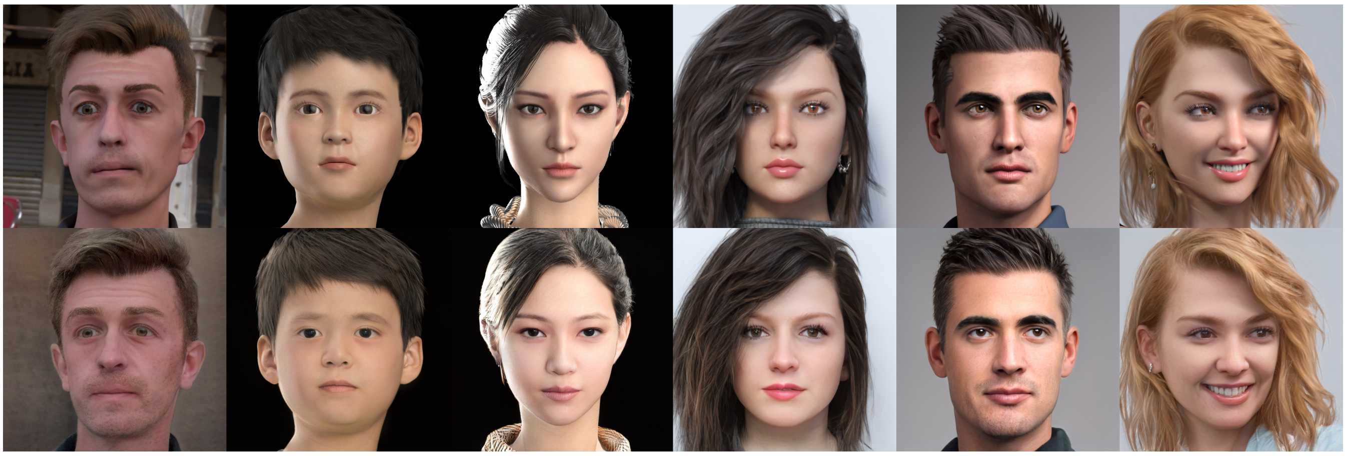


Figure 1: *Given rendered 3D faces as input (top row), our method effectively mitigates the "uncanny valley" effect and improves the overall authenticity of rendered portraits while preserving facial identity (bottom row). Please zoom in for a better view.*

Abstract

Despite rapid advances in computer graphics, creating high-quality photo-realistic virtual portraits is prohibitively expensive. Furthermore, the well-known "uncanny valley" effect in rendered portraits has a significant impact on the user experience, especially when the depiction closely resembles a human likeness, where any minor artifacts can evoke feelings of eeriness and repulsiveness. In this paper, we present a novel photo-realistic portrait generation framework that can effectively mitigate the "uncanny valley" effect and improve the overall authenticity of rendered portraits. 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), that is specifically designed for rendering-style portraits. We leverage this dataset to fine-tune the StyleGAN2 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 and ablation studies show the advantages of our method compared to state-of-the-art approaches.*

CCS Concepts

• ***Computing methodologies*** → ***Image processing;***

1. Introduction

Generating photo-realistic and indistinguishable faces from 3D renderings has long been a challenge. Over the last two decades,

[†] Xiaogang Jin is the corresponding author. E-mail: jin@cad.zju.edu.cn

the growth of entertainment industries such as animation, film, and video games has led to tremendous advances in high-quality face modeling and rendering technology. Under passive illumination, approaches based on multi-view stereo systems [BBB*10, BHB*11,BHPS10,FNH*17,RGB*20] can reconstruct high-quality face geometry. Following the pioneering work of Debevec et al. [DHT*00], a series of light-stage-based facial appearance capture methods [SXZ*20, GFT*11, GFT*15, GSSM15] have been proposed to capture the pore-level properties of a human face. While these approaches can help produce high-quality faces, they are highly expensive and time-consuming. Furthermore, even though these faces are of superior quality, they contain subtle unrealistic details that are immediately noticeable because humans are innately sensitive to such details when perceiving faces. The "uncanny valley" effect, first described by Japanese roboticist Masahiro Mori [Mor70], shows how imperfectly human-like objects, such as robots, 3D animations, and life-like dolls, can have a negative impact on user experience and interaction [Moo12]. Because these impersonations do not have a lifelike appearance, they can cause a sudden shift in a person's response from empathy to eerie, frightening, or revulsion, also known as "uncanny" sensations.

The rise of deep learning, in particular Generative Adversarial Networks (GANs) [GPAM*14], has inspired researchers to develop high-quality face generation methods [ZPIE17, CUYH20]. In recent years, StyleGAN [KLA19] and its variants [KLA*20, KAH*20,KAL*21] have paved the way for the semantic manipulation of photo-realistic portraits. The existing methods that can improve the realism of avatar faces [GKJS20,CWZ*21] are all based on projecting the rendering-style faces into the pretrained StyleGAN2 generator, thanks to its high generation quality and diversity. Garbin et al. [GKJS20] matches a non-photorealistic portrait to a latent code of the pretrained StyleGAN2 generator while maintaining pose, expression, hair, and lighting consistency. Despite the attempt to adapt to the real face domain, their method necessitates intricate and time-consuming processing. 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. [CWZ*21] 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 rendered portraits. 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 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 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 [KAH*20], we use *DRFHQ* to fine-tune the StyleGAN2 generator, which has been initialized with the weights of the StyleGAN2-*FFHQ* generator, resulting in a rendering-style StyleGAN2-*DRFHQ* generator. During fine-tuning, 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 rendered portraits.

In summary, our work makes the following contributions:

- We present a novel portrait generation framework to overcome the "uncanny valley" effect for rendered 3D faces.
- Based on a new high-quality rendering-style portrait dataset (*DRFHQ*), we propose a novel transfer learning approach to correlate portraits with different styles in the learned latent space while preserving facial identity.

2. Related Work

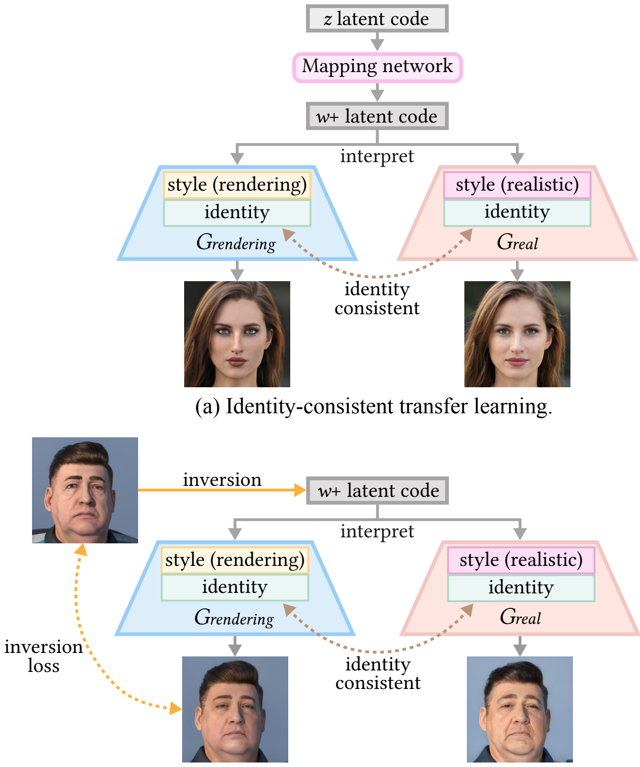
Portrait Synthesis. Human face modeling and rendering is a crucial and active research topic for applications in the entertainment, film, and television industries. Most physically-based rendering methods require a multi-view stereo system to reconstruct pore-level geometry and skin reflectance properties [FNH*17, LCC*22, LBZ*20, BHB*11, RGB*20]. To capture detailed human faces, a number of light-stage-based approaches [SXZ*20, GFT*11, GFT*15, GSSM15] have been developed based on the seminal work for facial appearance capturing and reconstruction [DHT*00]. 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 [SN07]. Researchers have suggested methods to measure the "uncanny valley" effect [SN07, HM17], however, it is difficult to eliminate such an unpleasant effect in traditional rendering. Since their introduction in 2020, neural radiance fields (NeRF) [MST*20] have spawned a slew of downstream applications, including face synthesis [GZX*22, HPX*22, ZZSC22]. Researchers also combine NeRF with generators [SLNG20,GLWT22, CMK*21, XLSL22] to support view-consistent image synthesis without the need for multi-view images of a specific person. However, existing face modelling and rendering methods still struggle to produce photo-realistic results that can avoid the "uncanny valley" effect. The introduction of generative adversarial networks (GANs) [GPAM*14] sparks an increasing number of face synthesis models [GAA*17, KALL18]. Among these works, StyleGAN [KLA19, KLA*20, KAL*21] is mostly favored due to its synthesis quality and manipulation ability, and serves as an inspiration for many downstream works [PDKR22,ZBG21]. We also build our framework based on the StyleGAN model since it not only provides important prior information of facial identity and appearance on various human faces, but also allows efficient portrait editing at a high level.

Face Style Transfer using StyleGAN. Portrait style transfer using StyleGAN is also related to our work. Pinkney and Adler [PA20] use a resolution-dependent method to interpolate different styles at appearance and geometry levels. StyleCariGAN [JJJ*21] modulates coarse layer feature maps of StyleGAN by shape exaggeration blocks to produce desirable caricature shape exaggerations. However, it requires a dataset that contains thousands of images, whereas other approaches have been proposed to reduce the dataset size to a few hundred [YJLL22a], $\sim$ 100 [SLL*21, MYC*22], $\sim$ 10 [OLL*21], or even to achieve one-shot domain adaptation [ZAFW22, ZLH*22]. Wu et al. [WNSL21] 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 [GPM*22] 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 [YJLL22b] to achieve consistent results in a sequence. Sang et al. [SZS*22] 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.

Face Realism Improvement using StyleGAN. Improving the realism of rendered faces is still a challenging issue. Garbin et al. [GKJS20] propose a zero-shot image projection algorithm that requires no training data to find the latent code that most closely matches the input rendered face. Their objective is the most similar to ours. However, their method requires a significant amount of processing time and may result in inconsistencies in facial identity. Chandran at al. [CWZ*21] use a multi-frame consistent method to project the traditional incomplete face rendering results into latent space to achieve photo-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 authenticity of resultant faces. The StyleGAN encoders [TAN*21,RAP*21,RMBCO22,APC21,ATM*22] and some optimization-based methods [AQW19, RMBCO22, AQW20] 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 3D faces while preserving the facial identity.

3. Method

Our objective is to improve the authenticity of digital 3D faces by substituting them with photo-realistic versions that are indistinguishable, all while preserving the avatar's inherent facial identity. 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 inbetween. 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 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 photo-realistic portrait with the identity of the input avatar.



(a) Identity-consistent transfer learning.

(b) The inference phase.

Figure 2: *The central idea of our method. (a) In identity-consistent transfer learning, a single latent code in the $W+$ latent space can be interpreted as portraits with the same facial identity but different image styles by G_{real} and $G_{rendering}$. (b) In our inference phase, we invert the input rendered portrait to the $W+$ latent space of $G_{rendering}$. The resulting latent codes can be interpreted as a realistic-style portrait by G_{real} while preserving the facial identity of the input rendered portrait. The rendered portrait is from the Diverse Human Faces [AI22] dataset.*

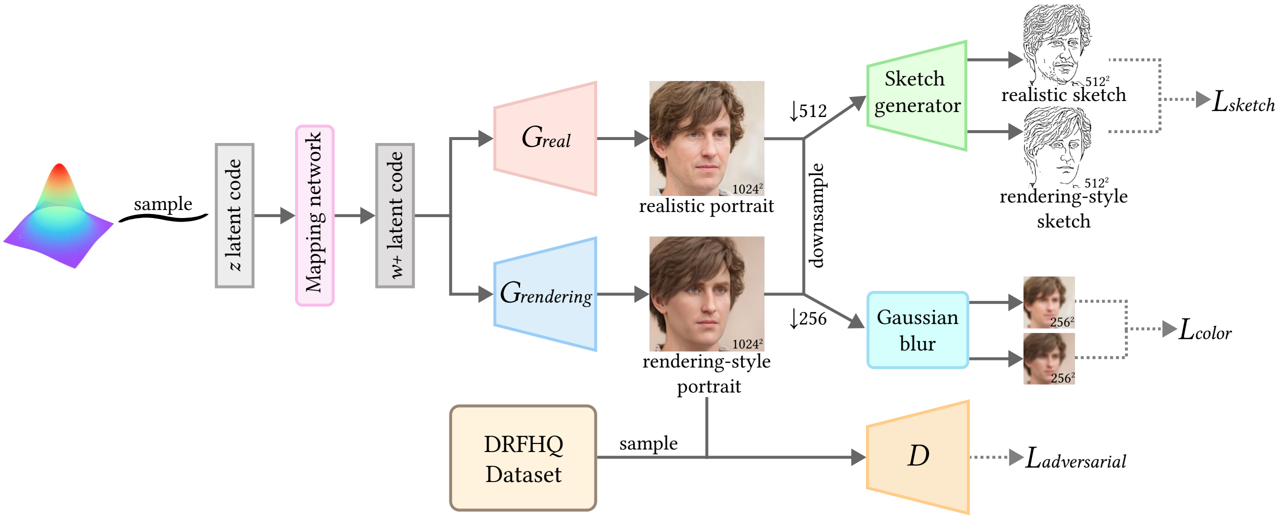


Figure 3: *An overview of our identity-consistent transfer learning network's training stage. We begin by initializing a rendering style generator $G_{rendering}$ with the weights of the pretrained StyleGAN2-FFHQ generator (G_{real}), which can provide various face priors. During the fine-tuning of the generator, we only update $G_{rendering}$ and the discriminator, while the sketch generator, G_{real}, and the mapping network are frozen. The L_{color} and L_{sketch} loss encourages small variations in color and face contour between the image generated by G_{real} and the image generated by $G_{rendering}$.*

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 rendered portraits 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 [Pro23]. *DRFHQ* contains 11,399 high-quality PNG images in 1024 × 1024 resolution, with a wide range of gender, age, pose, race, hairstyle, etc. We first align and crop the raw images using Dlib [KS14] 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 [WBH*21, LLQ*21, OAA20, AI22], their face resolution is insufficient for high-quality digital face display [WBH*21, AI22], or they only contain a small number of rendered faces [LLQ*21, OAA20], or they are rendered using a small number of face models (100 different identities) [AI22]. *DRFHQ* is the first high-quality rendering-style dataset with a face region resolution of 1024 × 1024 that can be extended to downstream tasks, to the best of our knowledge.

3.2. Identity-Consistent Transfer Learning

Inspired by StyleGAN-ada [KAH*20], 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. 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 (training) 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.

Sketch loss. Inspired by DeepFaceEditing [CLL*21], 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 [CLL*21] model, and $\downarrow_{512}$ denotes the interpolation operation that downsamples the images to 512 × 512. According to Eq. 1, the output of G_{real} and $G_{rendering}$ are fed into S separately to obtain

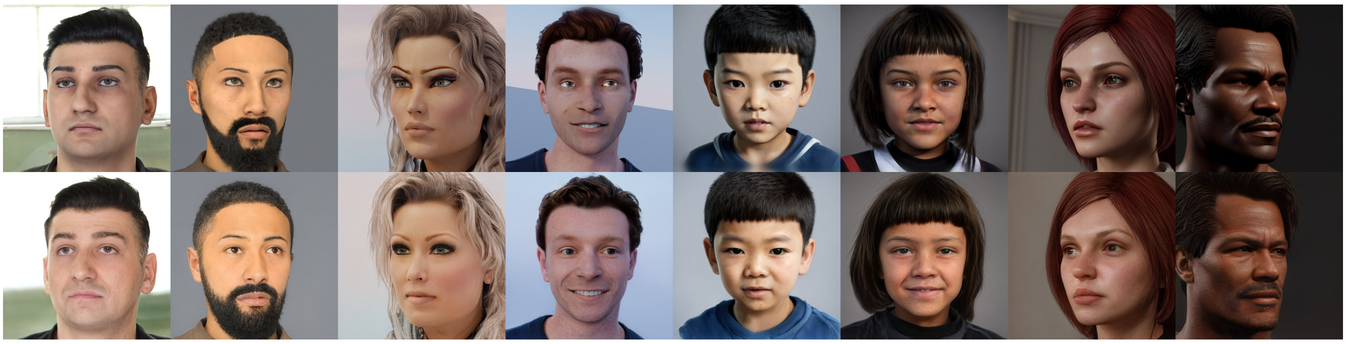


Figure 4: *Diverse photo-real faces generated by our method. The generated faces overcome the "uncanny valley" effect by maintaining identity, facial contour, and color. The input rendered portraits are from the Diverse Human Faces dataset (columns 1, 2), Paul Schultz at Flickr [Fli22] (columns 3, 4), and rendering-style images generated using Stable Diffusion [RBL*22] (columns 5-8).*

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 [ZIE*18]. However, LPIPS captures the facial appearance similarity, including texture and style-related details, preventing the generator from learning rendering-style. Inspired by [GKJS20], we solve this problem by removing the appearance details from the images. Specifically, we first downsample the images to 256 × 256 and apply Gaussian blur, then feed the images into the VGG16 network to compute the LPIPS loss:

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

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 × 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 a direct latent optimization [KLA*20] inversion method to project the rendered portrait x into the latent space of $G_{rendering}$. As we aim for the least distortion instead of the best editability, we optimize in the $W+$ latent space, which has greater expressive potential:

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

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.

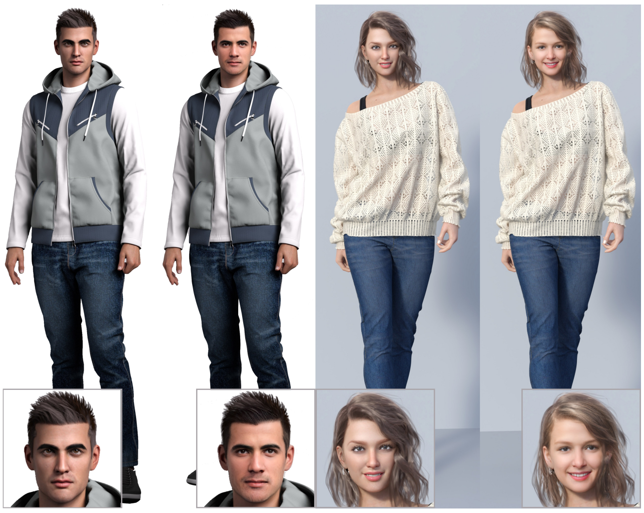


Figure 5: *We apply our method to digital apparel sample display images. Input images are courtesy of Yayat Punching at the CONNECT store [CLO22].*

4. Results

This section showcases the outcomes of our photo-realistic portrait generation framework. We present the results of our approach as applied to a series of rendering-style portraits. In Figs. 1 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 [AI22] , internet images, and rendering-style images generated using Stable Diffusion [RBL*22]. Additionally, we also showcase some examples where we stitch the generated realistic faces back onto the original apparel display renderings (refer to Fig. 5 and Fig. 16). Our generated realistic faces can easily blend in with the rendered garment and virtual avatar bodies with only minor post-processing (see Sec. 8). 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. 7) 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.

5. Experiments

5.1. Implementation Details

Networks. We use the StyleGAN2-ada architecture [KAH*20] 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 provided in the stylegan2 config of StyleGAN2-ada to fine-tune StyleGAN2-*DRFHQ* 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 fine-tuning when the discriminator had seen a total of 40k real images. PyTorch [PGM*19] 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 × 1024. 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 the *Diverse Human Faces* [AI22] dataset, the CONNECT store [CLO22] and rendering-style images generated using Stable Diffusion [RBL*22]. Specifically, we employ the fine-tuned and LoRA models based on Stable Diffusion 1.5 from `https://civitai.com/` for generating rendering-style images.

5.2. Comparison with State-of-the-Art Methods

In this section, 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, such as [GKJS20], [CWZ*21], 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. 5.2.1) and SDEdit (Sec. 5.2.2). Subsequently, we provide comparisons between our identity-consistent style-transfer method and SOTA style-transfer methods (Sec. 5.2.3).

5.2.1. Comparison with StyleGAN inversion methods

In this section, 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. 7, we use e4e [TAN*21], pSp [RAP*21], HyperStyle [ATM*22], ReStyle [APC21], and latent code optimization [RMBCO22], 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.

Quantitative evaluation. To the best of our knowledge, there is no currently 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. 5.2.1) 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.

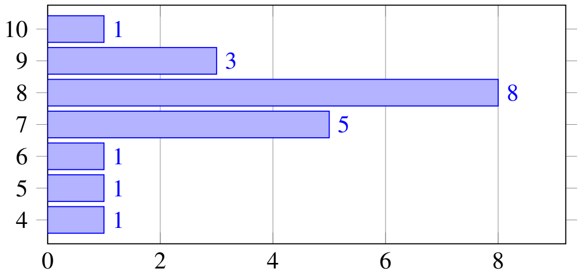


Figure 6: *The distribution of the user study on the authenticity comparison of the methods for improving facial realism. The y−axis shows the number of output portraits from our method chosen by participants (out of 10 sets), and the x−axis shows the number of participants. The results show that our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

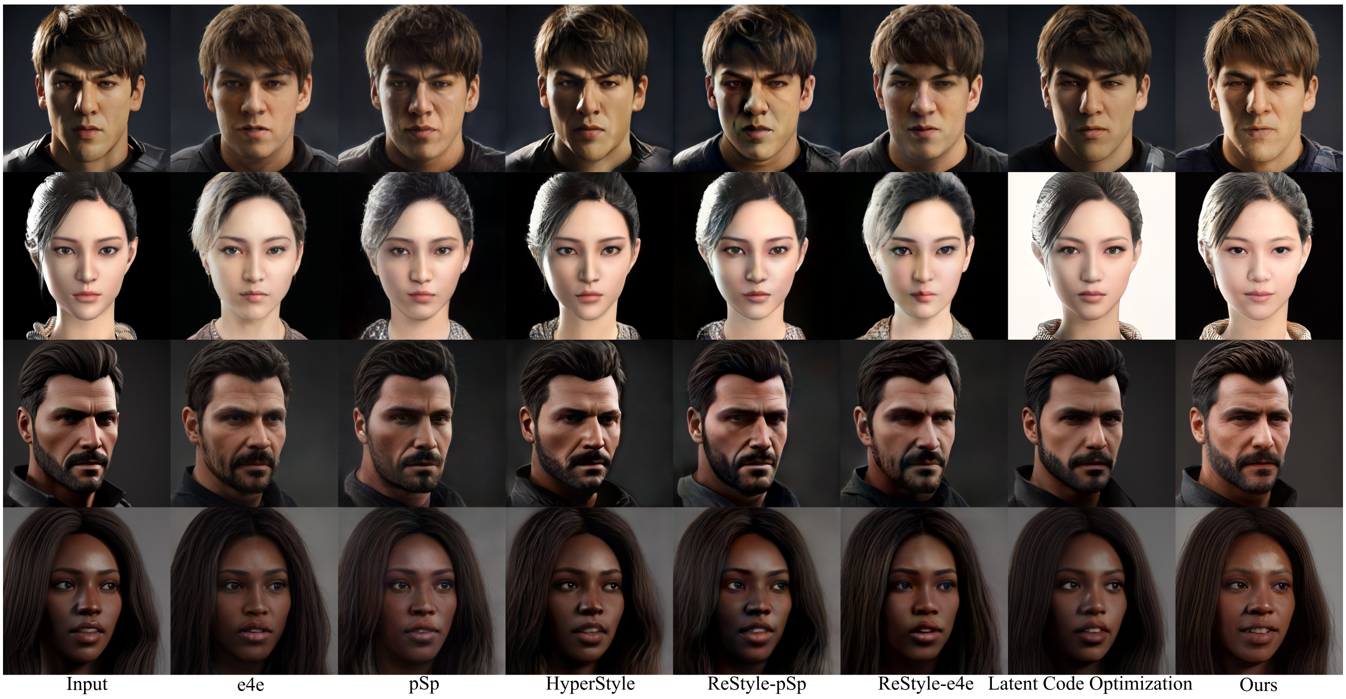


Figure 7: *Qualitative comparisons with state-of-the-art StyleGAN inversion methods. From left to right, we show the input image, the results of e4e, pSp, HyperStyle, ReStyle-pSp, ReStyle-e4e, latent code optimization, and ours. The testing input images (except for row 2) are generated by Stable Diffusion [RBL*22] model. Please zoom in for a better view.*

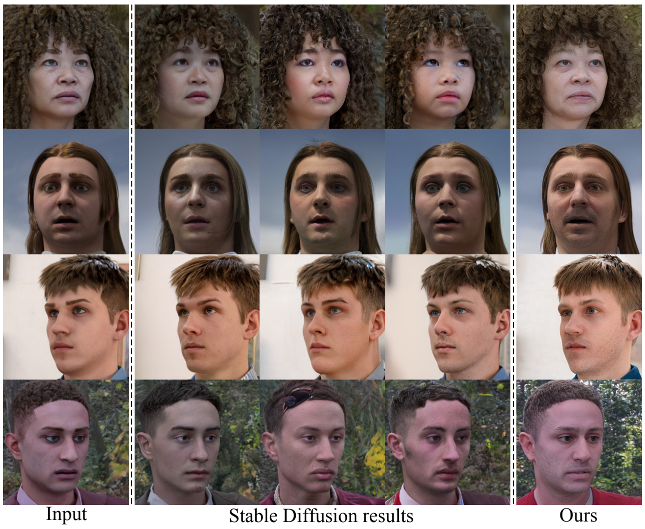


Figure 8: *Qualitative comparison with SDEdit [MHS*22]. From left to right, we present the input image, three randomly generated images by SDEdit, and ours. The testing input images are from the Diverse Human Faces dataset. Please zoom in for a better view.*

5.2.2. Comparison with SDEdit.

We also conduct a comparison with state-of-the-art diffusion-based method, SDEdit [MHS*22]. SDEdit projects an unrealistic image onto the manifold of natural images by adding noise and then removing it.

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 [RBL*22] trained on the *FFHQ* dataset at a resolution of 256×256. As illustrated in Fig. 8, the results produced by SDEdit do not ensure the complete removal of the rendering-style from the input (rows 2, 4), 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.

5.2.3. Style Transfer

In this section, we conduct qualitative and quantitative experiments to demonstrate 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 [ZAFW22, ZLH*22] 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 [GPM*22] and AgileGAN [SLL*21].

Qualitative evaluation. We present a comparison between the style transfer results of our identity-consistent style transfer method and those of StyleGAN-NADA [GPM*22] and AgileGAN [SLL*21] in Fig. 9. 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 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.

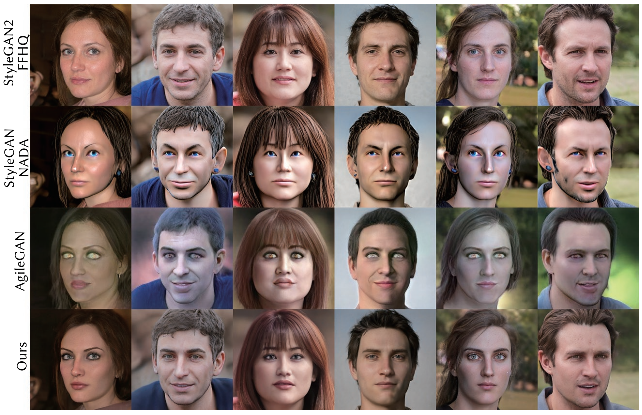


Figure 9: *Qualitative comparisons with state-of-the-art style transfer methods based on StyleGAN. From up to bottom, we present the images generated by StyleGAN2-FFHQ, StyleGAN-NADA, AgileGAN, and ours. Images in the same column are generated by the same latent code.*

Quantitative evaluation. To evaluate the performance in transferring style to that of *DRFHQ* dataset, we utilize Fréchet Inception Distance (FID) [HRU*17] to measure the overall similarity between the distribution of synthesized images and that of the *DRFHQ* dataset (see the 2nd column in Table 1). 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 the 3rd column in Table 1). In Table 1, StyleGAN2-*DRFHQ* represents our identity-consistent model fine-tuned on our *DRFHQ* dataset, AgileGAN-*DRFHQ* represents AgileGAN [SLL*21] fine-tuned on our *DRFHQ* dataset. Since StyleGAN-NADA [GPM*22] is text-guided and not trained on our *DRFHQ* dataset, FID is for reference only. 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.

Table 1: *Fréchet Inception Distances (FID) score for different StyleGAN-based style transfer methods and datasets, computed from randomly generated 50k images. Lower scores are better.*

Algorithm	*DRFHQ*	*FFHQ*
StyleGAN2-*DRFHQ* (Ours)	**24.5**	**16.3**
AgileGAN-*DRFHQ*	62.5	83.5
StyleGAN-NADA	49.9	53.8

To further assess the performance in facial identity preservation, we utilize a pretrained CurricularFace network [HWT*20] 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 CurricularFace network to measure facial identity. As shown in Table 2, our StyleGAN2-*DRFHQ* model exhibits superior performance in preserving facial identity during the process of style transfer.

Table 2: *Identity similarity measurement for SOTA StyleGAN-based style transfer methods, computed from randomly generated 2k images. Higher scores are better.*

Algorithm	Identity Similarity ↑
AgileGAN-*DRFHQ*	0.14
StyleGAN-NADA	0.34
StyleGAN2-*DRFHQ* (Ours)	**0.57**

5.3. Ablation Studies

In this section, we perform four ablation studies to validate the effectiveness of different components of our work. We first evaluate the two proposed losses (Sec. 5.3.1), then our transfer-learning-based framework (Sec. 5.3.2), the employed inversion method (Sec. 5.3.3), and finally our new high-quality rendering-style portrait dataset (Sec. 5.3.4).

5.3.1. Losses

We define StyleGAN2-*FFHQ* generator fine-tuned on our *DRFHQ* dataset without sketch and color constraints as the baseline. As shown in Fig. 10, we feed the same latent codes into various generator variants and compare the resulting portraits.

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.

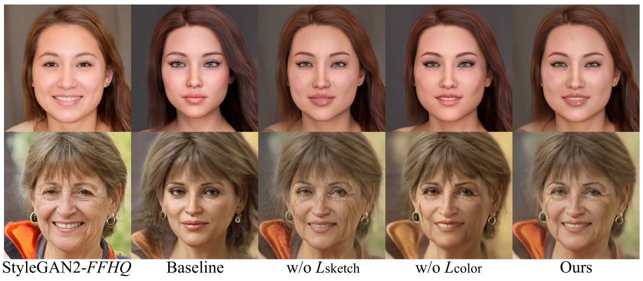


Figure 10: *Exemplars of the ablation study of the baseline network and ours. From left to right, we present the images generated by StyleGAN2-FFHQ, the baseline, the generator trained without L_{sketch}, the generator trained without L_{color}, and our full generator, respectively.*

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*.

5.3.2. 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.

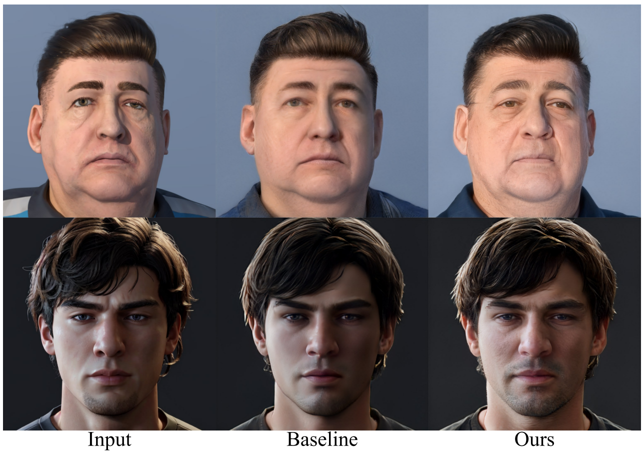


Figure 11: *Exemplars of the ablation study of the baseline method and ours. From left to right, we present the input rendered image (row 1 from the Diverse Human Faces dataset, row 2 generated by Stable Diffusion model), the result generated by the baseline method, and the result generated by ours.*

As shown in Fig. 11, 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}.

5.3.3. Inversion

In our framework, we use the latent code optimization described by Roich et al. [RMBCO22] as our inversion method during inference. We compare it to the following cutting-edge inversion approaches: e4e [TAN*21], ReStyle scheme on e4e (ReStyle-e4e) [APC21], and II2S [AQW19].

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. 12, 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.

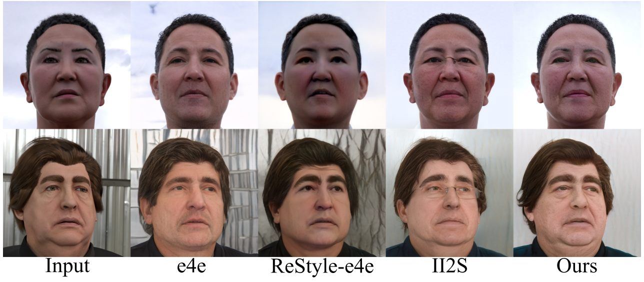


Figure 12: *Exemplars of the ablation study of different inversion methods. From left to right, we present the input rendered image (from the Diverse Human Faces dataset), the results generated using e4e, ReStyle-e4e, II2S, and ours as the inversion method.*

5.3.4. 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 [AI22]. To this end, we replace our *DRFHQ* dataset with *Diverse Human Faces* dataset during generator fine-tuning, while maintaining method consistency.

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. 13, 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.

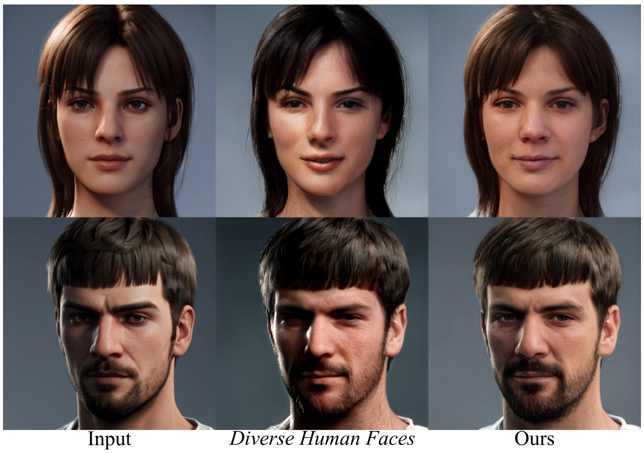


Figure 13: *Exemplars of the ablation study of the Diverse Human Faces dataset and our DRFHQ dataset. From left to right, we present the input rendered image, the results generated using the Diverse Human Faces dataset-based framework, and ours. The input rendered images are generated by Stable Diffusion [RBL*22] 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 [ZIE*18] and L2 loss to assess dataset performance in information preservation. Table 3 demonstrates that our *DRFHQ* dataset outperforms the *Diverse Human Faces* dataset in both metrics, indicating superior overall information preservation.

Table 3: *Mean LPIPS and L2 losses from 150 pairs of images for the Diverse Human Faces dataset-based and our DRFHQ dataset-based frameworks. Lower values indicate better performance.*

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

6. Limitations and Future Work.

Our method has some limitations as shown in Fig. 14. When the input faces contain accessories such as unique beards, glasses, hats, and headsets, the faces generated by our model have visible inconsistencies with the original images. 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.

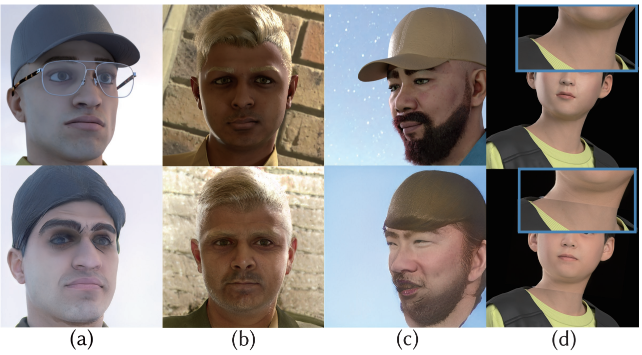


Figure 14: *Example of failure cases. Our method may fail in cases of faces with glasses and hats (a), complicated background (b), and large posture (c). There exist chromatic aberration and misalignment when we paste the resulting image onto the full-body apparel sample display images (d). The rendered portraits in (a, b, c) are from Diverse Human Faces [AI22] dataset.*

Our method meets the challenges to reconstruct the background of images. We attribute this to StyleGAN's weak expressive capacity for complicated backgrounds. This limitation can be solved by removing the generated background using the alpha matte.

We notice that our approach cannot process those faces with extreme poses. 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. To achieve seamless integration, a lightweight post-processing of the resulting portrait is further applied (see Section 8).

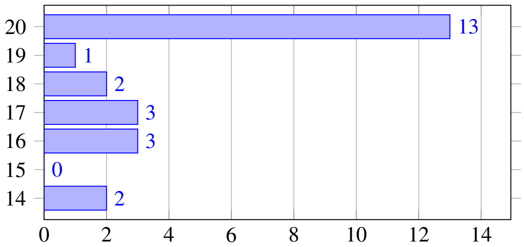


Figure 15: *Distribution of the user study on the authenticity of the full-body apparel display images. The y−axis represents the number of output images from our method selected by the participants (out of 20 pairs), and the x−axis represents the number of participants. Results demonstrate that stitching the resulting realistic faces back onto the full-body apparel display images can effectively improve the overall authenticity.*

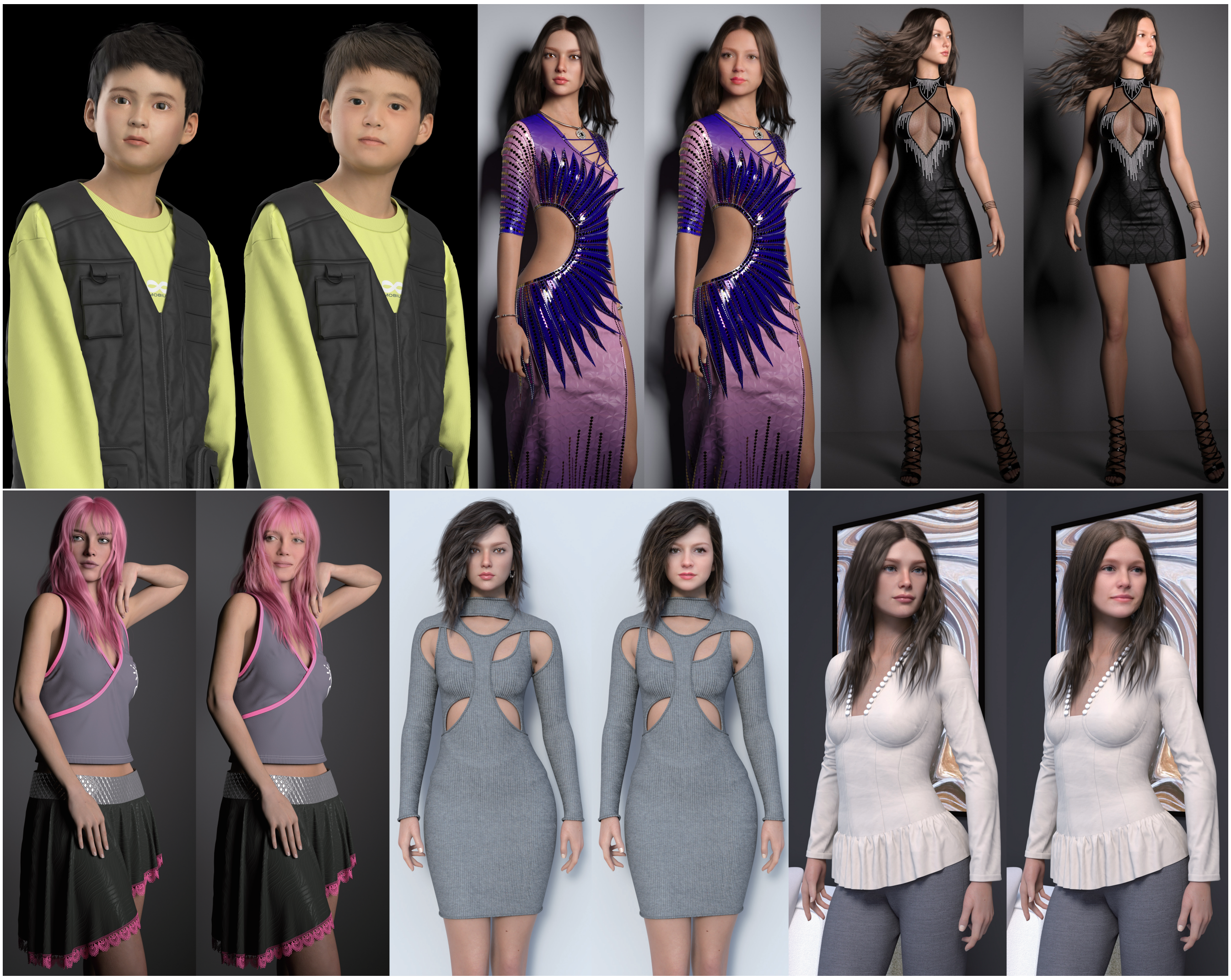


Figure 16: *Our method's application in digital sample display. We replace original rendered 3D avatars' faces with photo-realistic faces generated by our method. The results show that the generated photo-real faces blend in with the rendered garments and virtual avatar bodies, effectively increasing the authenticity of the digital apparel sample display images. The input images are courtesy of Yayat Punching at the CONNECT store [CLO22], except for the first one in row 1.*

7. Application in Digital Sample Display

Our proposed method can also be applied to improve the authenticity of digital sample display images. Fig. 16 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 [CLO22].

To further validate the improvement in the authenticity of digital sample display, we collected 20 full-body apparel display images and processed them using our framework, yielding 20 pairs of images with faces of rendering-style and realistic-style, respectively. We present these 20 pairs of test cases in sequence to 24 participants, with the original and processed images in each pair randomly displayed in position for authenticity comparison. Fig. 15 demonstrated that the vast majority of the full-body apparel display images replaced faces by our method are considered more realistic. This validates that stitching the resulting photo-realistic faces back onto the full-body apparel display images can effectively improve the overall authenticity.

8. Lightweight Post-Processing

As mentioned above, one of the applications of our method is to enhance the authenticity of digital apparel display images. However, as shown in Fig. 14 (d) and Fig. 17, directly pasting the resulting portrait back onto the original rendered digital apparel display image may lead to small chromatic aberration and misalignment.

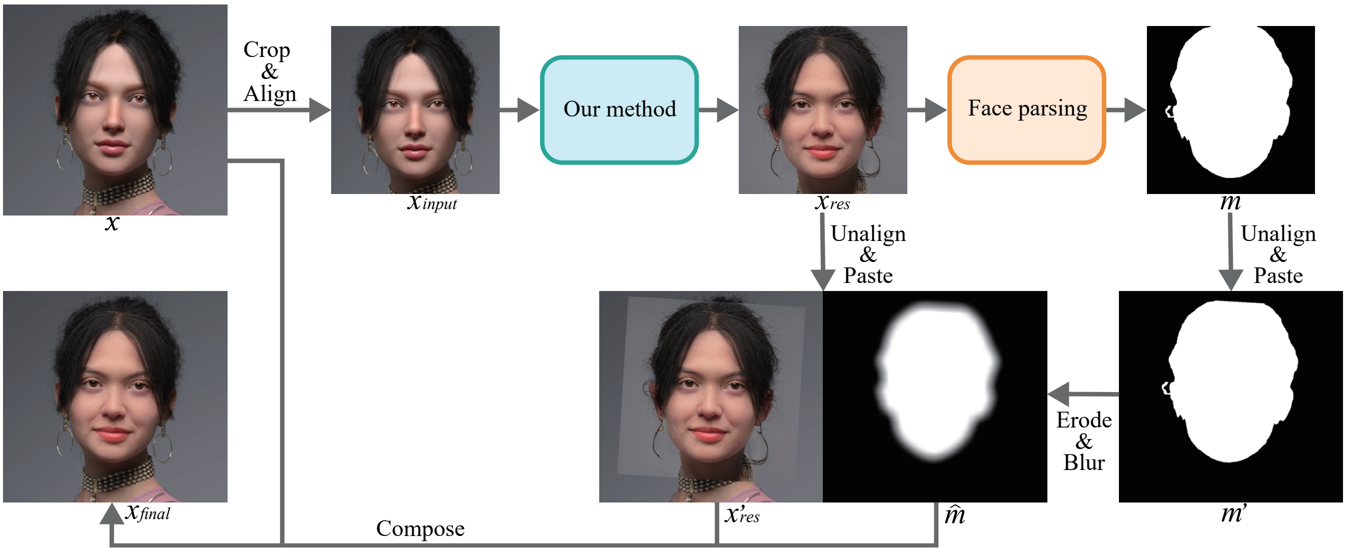


Figure 17: *An overview of our lightweight post-processing method.*

To address this issue, we propose a lightweight post-processing method.

Specifically, we apply face parsing [zll19] 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:

$$x_{final} = \hat{m} \odot x'_{res} + (1 - \hat{m}) \odot x, \tag{5}$$

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

9. Conclusions

We present a novel identity-consistent transfer learning method that can effectively 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 apply our method to digital apparel sample display, and experiments show that it can improve the overall realism of digital apparel samples. Moreover, our rendering-style *DRFHQ* dataset has the potential to motivate other applications such as virtual avatar synthesis and editing.

References

[AI22] AI S.: Synthesis ai. Website, 2022. https://opensynthetics.com/dataset/diverse-human-faces-dataset/. 3, 4, 5, 6, 9, 10

[APC21] ALALUF Y., PATASHNIK O., COHEN-OR D.: Restyle: A residual-based stylegan encoder via iterative refinement. In *IEEE/CVF International Conference on Computer Vision, ICCV* (2021), IEEE, pp. 6691–6700. 3, 6, 9

[AQW19] ABDAL R., QIN Y., WONKA P.: Image2stylegan: How to embed images into the stylegan latent space? In *IEEE/CVF International Conference on Computer Vision, ICCV* (2019), IEEE, pp. 4431–4440. 3, 9

[AQW20] ABDAL R., QIN Y., WONKA P.: Image2stylegan++: How to edit the embedded images? In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2020), Computer Vision Foundation / IEEE, pp. 8293–8302. 3

[ATM*22] ALALUF Y., TOV O., MOKADY R., GAL R., BERMANO A.: Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2022), IEEE, pp. 18490–18500. 3, 6

[BBB*10] BEELER T., BICKEL B., BEARDSLEY P., SUMNER B., GROSS M.: High-quality single-shot capture of facial geometry. SIGGRAPH '10, Association for Computing Machinery. 2

[BHB*11] BEELER T., HAHN F., BRADLEY D., BICKEL B., BEARDSLEY P., GOTSMAN C., SUMNER R. W., GROSS M.: High-quality passive facial performance capture using anchor frames. *ACM Trans. Graph. 30*, 4 (2011), 75:1–75:10. 2

[BHPS10] BRADLEY D., HEIDRICH W., POPA T., SHEFFER A.: High resolution passive facial performance capture. In *ACM SIGGRAPH 2010 Papers* (New York, NY, USA, 2010), SIGGRAPH '10, Association for Computing Machinery. 2

[CLL*21] CHEN S.-Y., LIU F.-L., LAI Y.-K., ROSIN P. L., LI C., FU H., GAO L.: Deepfaceediting: Deep face generation and editing with disentangled geometry and appearance control. *ACM Trans. Graph. 40*, 4 (2021), 90:1–15. 4

[CLO22] CLO: Clo virtual fashion inc. Website, 2022. connect.clo-set.com. 5, 6, 11

[CMK*21] CHAN E. R., MONTEIRO M., KELLNHOFER P., WU J., WETZSTEIN G.: Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2021), pp. 5799–5809. 2

[CUYH20] CHOI Y., UH Y., YOO J., HA J.-W.: Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020). 2

[CWZ*21] CHANDRAN P., WINBERG S., ZOSS G., RIVIERE J., GROSS M. H., GOTARDO P., BRADLEY D.: Rendering with style: combining traditional and neural approaches for high-quality face rendering. *ACM Trans. Graph. 40*, 6 (2021), 223:1–223:14. 2, 3, 6

[DHT*00] DEBEVEC P., HAWKINS T., TCHOU C., DUIKER H.-P., SAROKIN W., SAGAR M.: Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (USA, 2000), SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., p. 145–156. 2

[Fli22] FLICKR: Flickr. Website, 2022. flickr.com. 5

[FNH*17] FYFFE G., NAGANO K., HUYNH L., SAITO S., BUSCH J., JONES A., LI H., DEBEVEC P. E.: Multi-view stereo on consistent face topology. *Comput. Graph. Forum 36*, 2 (2017), 295–309. 2

[GAA*17] GULRAJANI I., AHMED F., ARJOVSKY M., DUMOULIN V., COURVILLE A. C.: Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems* (2017), pp. 5767–5777. 2

[GFT*11] GHOSH A., FYFFE G., TUNWATTANAPONG B., BUSCH J., YU X., DEBEVEC P.: Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph. 30*, 6 (2011), 1–10. 2

[GFT*15] GRAHAM P., FYFFE G., TONWATTANAPONG B., GHOSH A., DEBEVEC P.: Near-instant capture of high-resolution facial geometry and reflectance. In *ACM SIGGRAPH 2015 Talks* (2015), SIGGRAPH '15, Association for Computing Machinery, p. 32:1. 2

[GKJS20] GARBIN S. J., KOWALSKI M., JOHNSON M., SHOTTON J.: High resolution zero-shot domain adaptation of synthetically rendered face images. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVIII* (2020), vol. 12373 of *Lecture Notes in Computer Science*, Springer, pp. 220–236. 2, 3, 5, 6

[GLWT22] GU J., LIU L., WANG P., THEOBALT C.: Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *10th International Conference on Learning Representations, ICLR* (2022), OpenReview.net, pp. 1–13. 2

[GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *Advances in Neural Information Processing Systems* (2014), vol. 27, Curran Associates, Inc. 2

[GPM*22] GAL R., PATASHNIK O., MARON H., BERMANO A. H., CHECHIK G., COHEN-OR D.: Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph. 41*, 4 (2022), 141:1–141:13. 3, 8

[GSSM15] GOTARDO P., SIMON T., SHEIKH Y., MATTHEWS I.: Photogeometric scene flow for high-detail dynamic 3d reconstruction. In *IEEE/CVF International Conference on Computer Vision, ICCV* (2015), IEEE Computer Society, pp. 846–854. 2

[GZX*22] GAO X., ZHONG C., XIANG J., HONG Y., GUO Y., ZHANG J.: Reconstructing personalized semantic facial nerf models from monocular video. *ACM Trans. Graph. 41*, 6 (2022), 200:1–200:12. 2

[HM17] HO C., MACDORMAN K. F.: Measuring the uncanny valley effect - refinements to indices for perceived humanness, attractiveness, and eeriness. *Int. J. Soc. Robotics 9*, 1 (2017), 129–139. 2

[HPX*22] HONG Y., PENG B., XIAO H., LIU L., ZHANG J.: Headnerf: A real-time nerf-based parametric head model. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (June 2022), pp. 20374–20384. 2

[HRU*17] 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 *Advances in Neural Information Processing Systems* (2017), Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R., (Eds.), vol. 30, Curran Associates, Inc. 8

[HWT*20] HUANG Y., WANG Y., TAI Y., LIU X., SHEN P., LI S., LI J., HUANG F.: Curricularface: Adaptive curriculum learning loss for deep face recognition, 2020. arXiv:2004.00288. 8

[JJJ*21] JANG W., JU G., JUNG Y., YANG J., TONG X., LEE S.: Stylecarigan: caricature generation via stylegan feature map modulation. *ACM Trans. Graph. 40*, 4 (2021), 116:1–116:16. 3

[KAH*20] KARRAS T., AITTALA M., HELLSTEN J., LAINE S., LEHTINEN J., AILA T.: Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems* (2020), vol. 33, Curran Associates, Inc., pp. 12104–12114. 2, 4, 6

[KAL*21] KARRAS T., AITTALA M., LAINE S., HÄRKÖNEN E., HELLSTEN J., LEHTINEN J., AILA T.: Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems* (2021), vol. 34, Curran Associates, Inc., pp. 852–863. 2

[KALL18] KARRAS T., AILA T., LAINE S., LEHTINEN J.: Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR* (2018). 2

[KLA19] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2019), pp. 4401–4410. 2

[KLA*20] KARRAS T., LAINE S., AITTALA M., HELLSTEN J., LEHTINEN J., AILA T.: Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2020), pp. 8107–8116. 2, 5

[KS14] KAZEMI V., SULLIVAN J.: One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2014), pp. 1867–1874. 4

[LBZ*20] LI R., BLADIN K., ZHAO Y., CHINARA C., INGRAHAM O., XIANG P., REN X., PRASAD P., KISHORE B., XING J., LI H.: Learning formation of physically-based face attributes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2020), pp. 3407–3416. 2

[LCC*22] LIU S., CAI Y., CHEN H., ZHOU Y., ZHAO Y.: Rapid face asset acquisition with recurrent feature alignment. *ACM Trans. Graph. 41*, 6 (2022), 214:1–214:17. 2

[LLQ*21] LIU M., LI Q., QIN Z., ZHANG G., WAN P., ZHENG W.: Blendgan: Implicitly gan blending for arbitrary stylized face generation. In *Advances in Neural Information Processing Systems* (2021), vol. 34, Curran Associates, Inc., pp. 29710–29722. 4

[MHS*22] MENG C., HE Y., SONG Y., SONG J., WU J., ZHU J.-Y., ERMON S.: Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. arXiv:2108.01073. 7

[Moo12] MOORE R. K.: A bayesian explanation of the 'uncanny valley'effect and related psychological phenomena. *Scientific reports 2*, 1 (2012), 1–5. 2

[Mor70] MORI M.: Bukimi no tani [the uncanny valley]. *Energy 7* (1970), 33–35. 2

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I* (2020), vol. 12346 of *Lecture Notes in Computer Science*, Springer, pp. 405–421. 2

[MYC*22] MEN Y., YAO Y., CUI M., LIAN Z., XIE X.: Dct-net: domain-calibrated translation for portrait stylization. *ACM Trans. Graph. 41*, 4 (2022), 140:1–140:9. 3

[OAA20] OLIVER M. M., AMENGUAL ALCOVER E.: Uibvfed: Virtual facial expression dataset. *Plos one 15*, 4 (2020), e0231266. 4

[OLL*21] OJHA U., LI Y., LU J., EFROS A. A., LEE Y. J., SHECHTMAN E., ZHANG R.: Few-shot image generation via cross-domain correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (June 2021), pp. 10743–10752. 3

[PA20] PINKNEY J. N. M., ADLER D.: Resolution dependent GAN interpolation for controllable image synthesis between domains. *CoRR abs/2010.05334* (2020). arXiv:2010.05334. 3

[PDKR22] PARIHAR R., DHIMAN A., KARMALI T., R V.: Everything is there in latent space: Attribute editing and attribute style manipulation by stylegan latent space exploration. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022* (2022), ACM, pp. 1828–1836. 2

[PGM*19] PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KÖPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J., CHINTALA S.: *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA, 2019. 6

[Pro23] PRODUCTIONS D.: Daz productions inc. Website, 2023. www.daz3d.com/gallery. 4

[RAP*21] RICHARDSON E., ALALUF Y., PATASHNIK O., NITZAN Y., AZAR Y., SHAPIRO S., COHEN-OR D.: Encoding in style: A stylegan encoder for image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2021), Computer Vision Foundation / IEEE, pp. 2287–2296. 3, 6

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), pp. 10684–10695. 5, 6, 7, 10

[RGB*20] RIVIERE J., GOTARDO P., BRADLEY D., GHOSH A., BEELER T.: Single-shot high-quality facial geometry and skin appearance capture. *ACM Trans. Graph. 39*, 4 (2020), 81:1–81:12. 2

[RMBCO22] ROICH D., MOKADY R., BERMANO A. H., COHEN-OR D.: Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph. 42*, 1 (2022), 6:1–6:13. 3, 6, 9

[SLL*21] SONG G., LUO L., LIU J., MA W., LAI C., ZHENG C., CHAM T.: Agilegan: stylizing portraits by inversion-consistent transfer learning. *ACM Trans. Graph. 40*, 4 (2021), 117:1–117:13. 3, 8

[SLNG20] SCHWARZ K., LIAO Y., NIEMEYER M., GEIGER A.: GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In *Advances in Neural Information Processing Systems* (2020), vol. 33, pp. 20154–20166. 2

[SN07] SEYAMA J., NAGAYAMA R. S.: The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence Teleoperators Virtual Environ. 16*, 4 (2007), 337–351. 2

[SXZ*20] SUN T., XU Z., ZHANG X., FANELLO S., RHEMANN C., DEBEVEC P., TSAI Y.-T., BARRON J. T., RAMAMOORTHI R.: Light stage super-resolution: Continuous high-frequency relighting. *ACM Trans. Graph. 39*, 6 (2020), 260:1–260:12. 2

[SZS*22] SANG S., ZHI T., SONG G., LIU M., LAI C., LIU J., WEN X., DAVIS J., LUO L.: Agileavatar: Stylized 3d avatar creation via cascaded domain bridging. In *SIGGRAPH Asia 2022 Conference Papers, SA 2022* (2022), ACM, pp. 23:1–23:8. 3

[TAN*21] TOV O., ALALUF Y., NITZAN Y., PATASHNIK O., COHEN-OR D.: Designing an encoder for stylegan image manipulation. *ACM Trans. Graph. 40*, 4 (2021), 133:1–133:14. 3, 6, 9

[WBH*21] WOOD E., BALTRUŠAITIS T., HEWITT C., DZIADZIO S., JOHNSON M., ESTELLERS V., CASHMAN T. J., SHOTTON J.: Fake it till you make it: Face analysis in the wild using synthetic data alone, 2021. arXiv:2109.15102. 4

[WNSL21] WU Z., NITZAN Y., SHECHTMAN E., LISCHINSKI D.: Stylealign: Analysis and applications of aligned stylegan models. *arXiv preprint arXiv:2110.11323* (2021). 3

[XLSL22] XUE Y., LI Y., SINGH K. K., LEE Y. J.: GIRAFFE HD: A high-resolution 3d-aware generative model. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2022), pp. 18419–18428. 2

[YJLL22a] YANG S., JIANG L., LIU Z., LOY C. C.: Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), pp. 7693–7702. 3

[YJLL22b] YANG S., JIANG L., LIU Z., LOY C. C.: Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Trans. Graph. 41*, 6 (2022), 203:1–203:15. 3

[ZAFW22] ZHU P., ABDAL R., FEMIANI J., WONKA P.: Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. In *10th International Conference on Learning Representations, ICLR* (2022), OpenReview.net, pp. 1–12. 3, 8

[ZBG21] ZHANG L., BAI X., GAO Y.: Sals-gan: Spatially-adaptive latent space in stylegan for real image embedding. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021* (2021), ACM, pp. 5176–5184. 2

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2018), Computer Vision Foundation / IEEE Computer Society, pp. 586–595. 5, 10

[ZLH*22] ZHANG Z., LIU Y., HAN C., GUO T., YAO T., MEI T.: Generalized one-shot domain adaptation of generative adversarial networks. In *Advances in Neural Information Processing Systems* (2022). 3, 8

[zll19] ZLLRUNNING: face-parsing.pytorch. https://github.com/zllrunning/face-parsing.PyTorch, 2019. 12

[ZPIE17] ZHU J.-Y., PARK T., ISOLA P., EFROS A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Oct 2017). 2

[ZZSC22] ZHUANG Y., ZHU H., SUN X., CAO X.: Mofanerf: Morphable facial neural radiance field. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III* (Berlin, Heidelberg, 2022), Springer-Verlag, p. 268–285. 2