AgileGAN: Stylizing Portraits by Inversion-Consistent Transfer Learning

GUOXIAN SONG*, Nanyang Technological University, Singapore
LINJIE LUO, JING LIU, WAN-CHUN MA, and CHUNPONG LAI, ByteDance Inc, USA
CHUANXIA ZHENG and TAT-JEN CHAM, Nanyang Technological University, Singapore

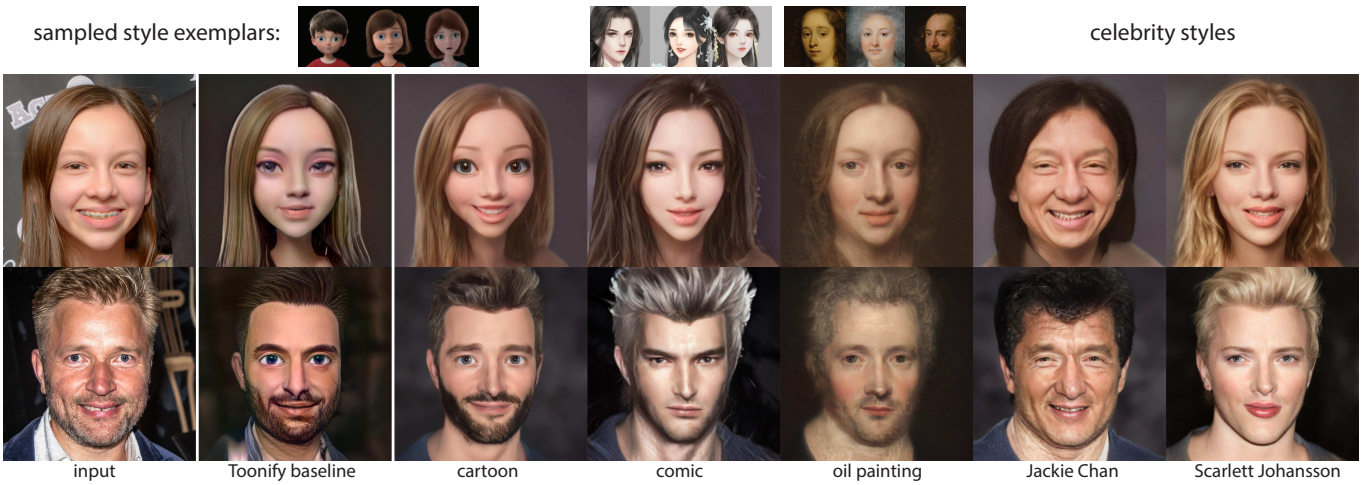

Fig. 1. Top row (small): some sampled style exemplars. Bottom two rows: input images, results from Toonify [Pinkney and Adler 2020], and our results for multiple styles. Given a single input image, our method can quickly (130 ms) and automatically generate high quality (1024×1024) portraits in various artistic styles. For a new style, our agile training strategy only requires ∼100 style exemplars and can be trained in 1 hour. Please magnify to see details. Cartoon style exemplars are from our dataset. Others are courtesy of Qingqian for comic style images; Cornelis Jonson van Ceulen the Elder,Abraham de Vries,Villers for Oil painting; The input images are courtesy of Michael Bull(Public Domain) and Ritvars Stankevičs(Public Domain).

Portraiture as an art form has evolved from realistic depiction into a plethora of creative styles. While substantial progress has been made in automated stylization, generating high quality stylistic portraits is still a challenge, and even the recent popular Toonify suffers from several artifacts when used on real input images. Such StyleGAN-based methods have focused on finding the best latent inversion mapping for reconstructing input images; however, our key insight is that this does not lead to good generalization to different portrait styles. Hence we propose AgileGAN, a framework that can generate high quality stylistic portraits via inversion-consistent transfer learning. We introduce a novel hierarchical variational autoencoder to ensure the inverse mapped distribution conforms to the original latent Gaussian distribution, while augmenting the original space to a multi-resolution latent space so as to better encode different levels of detail. To better capture attribute-dependent stylization of facial features, we also present an attribute-aware generator and adopt an early stopping strategy to avoid overfitting small training datasets. Our approach provides greater agility in creating high quality and high resolution (1024×1024) portrait stylization models, requiring only a limited number of style exemplars (∼100) and short training time (∼1 hour). We collected several style datasets for evaluation including 3D cartoons, comics, oil paintings and celebrities. We show that we can achieve superior portrait stylization quality to previous state-of-the-art methods, with comparisons done qualitatively, quantitatively and through a perceptual user study. We also demonstrate two applications of our method, image editing and motion retargeting.

CCS Concepts: • **Computing methodologies → Non-photorealistic rendering**.

Additional Key Words and Phrases: Portrait Generation, Stylization, Image-to-Image Translation, StyleGAN

ACM Reference Format:
Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. 2021. AgileGAN: Stylizing Portraits by Inversion-Consistent Transfer Learning. *ACM Trans. Graph.* 37, 4, Article 111 (August 2021), 13 pages. https://doi.org/10.1145/3450626.3459771

1 INTRODUCTION

Portraiture, the art of depicting the appearance of a subject, is an important art form dating back to the beginning of civilization. It has

*This work was done during Guoxian's internship at ByteDance Inc.

Authors' addresses: Guoxian Song, guoxian001@e.ntu.edu.sg, Nanyang Technological University, Singapore, Singapore; Linjie Luo, linjie.luo@bytedance.com; Jing Liu, jing.liu@bytedance.com; Wan-Chun Ma, wanchun.ma@bytedance.com; Chunpong Lai, chunpong.lai@bytedance.com, ByteDance Inc, California, San Jose, USA; Chuanxia Zheng, chuanxia001@e.ntu.edu.sg; Tat-Jen Cham, astjcham@ntu.edu.sg, Nanyang Technological University, Singapore, Singapore.

0730-0301/2021/8-ART111 $15.00
https://doi.org/10.1145/3450626.3459771

evolved beyond faithful depiction into more creative interpretations with a plethora of styles, such as abstract art, Cubism and cartoon.

Automatically stylized portraiture has undergone rapid progress in recent years due to advances in deep learning. Early methods involving neural style transfer [Gatys et al. 2016; Li and Wand 2016; Ruder et al. 2016] have convincingly demonstrated the ability to transfer textural styles from an exemplar source to target images, with real photos transformed into Van Gogh or Picasso paintings. However, when it comes to portraiture, these methods largely failed to capture the important geometry-dependent motifs of different portraiture styles, thus falling short in stylization quality.

Image-to-image translation methods were later introduced to "translate" images from a source domain to a target domain using paired datasets in a supervised manner [Isola et al. 2017; Wang et al. 2018], or using unpaired datasets in an unsupervised setting [Huang et al. 2018; Liu et al. 2017; Zhu et al. 2017]. These methods have been explored for portrait stylization, e.g. self-to-anime [Kim et al. 2020] and cartoon [Li 2018]. However, supervised approaches require paired datasets for training that would be manually onerous if not infeasible, while the unsupervised approaches not only need a large amount of unpaired data, but also often face difficulties with stable training convergence and in generating high-resolution results.

A recent portrait stylization pipeline, Toonify [Pinkney and Adler 2020], was proposed which builds on a pre-trained model of the high-resolution generative neural network StyleGAN2 [Karras et al. 2020b]. Using only around a few hundred unpaired exemplars, it had the ability to generate promising results in cartoon style, by employment of *transfer learning* to adapt StyleGAN2 to the given style exemplars. When given an input image, the corresponding latent code was obtained by an optimization-based *inversion* in one of the StyleGAN2 latent spaces, which is then used to generate the stylized output via the adapted StyleGAN2 model. Despite its strong generalization ability given only limited exemplars, the stylization of *real* input images (in contrast to StyleGAN2 realistically synthesized ones) nonetheless still resulted in various artifacts, likely due to the sub-optimality of the inversion method used.

Our key insight is that attempting to find the best inversion mapping in terms of reconstruction in the original StyleGAN2 is in fact misguided, because what is best for realistic images may not be the best for other stylized generators. What we discover instead is that if we learned an inversion mapping that also optimizes for matching the distribution of latent codes to the Gaussian latent distribution in the original StyleGAN2, the inversion mapping works better across a range of different stylized generators. In other words, *matching latent distributions when learning the inversion leads to the best robust embedding across different styles, and is better than aiming for the best reconstruction embedding for realistic images.* See Fig.2.

To this end, we propose *AgileGAN*, a novel *inversion-consistent transfer learning* framework for high quality portrait stylization using only limited exemplars. This allows us to *agilely* create high quality and high resolution portrait stylization models in a variety of target styles (Figure 1).

To achieve inversion consistency in our AgileGAN framework, we introduce a novel hierarchical Variational Autoencoder (hVAE) to perform inversion. Compared to recent latent space inversion methods [Abdal et al. 2019a; Karras et al. 2019; Tewari et al. 2020; Zhu et al. 2016] that usually operate on the less entangled latent space W, our hVAE approach ensures the mapping conforms to the multi-variate Gaussian distribution of the original StyleGAN2 latent space. Furthermore, our hVAE approach is hierarchical in the sense that we augmented the StyleGAN2's original Z latent space to a multi-resolution latent space Z^+ to better encode different levels of detail in the image. We show that using our Z^+ augmentation and hVAE can significantly improve stylization quality.

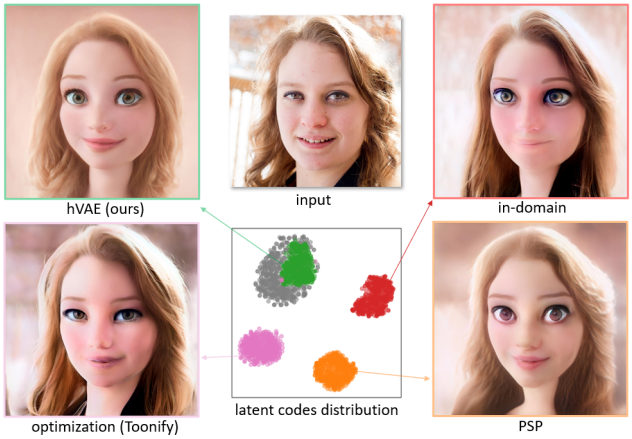

Fig. 2. t-SNE visualization of the latent code distributions for different inversion methods [Karras et al. 2020b; Richardson et al. 2020; Zhu et al. 2020], and the relation to stylized image quality. The gray dots are sampled from the original StyleGAN2 latent distribution. Having a latent code distribution that is better aligned to the original leads to more pleasant results. More details about the t-SNE distributions are provided in Sec.4.2. The input image is courtesy of Ritvars Stankevičs(Public Domain).

To improve the training efficiency with a high resolution dataset, we decompose the training process into two stages. First, we train hVAE for inversion encoding using the original StyleGAN2 as the decoder with fixed pre-trained weights. During training, we enforce reconstruction loss, user identity loss, perceptual loss and KL divergence loss for VAEs. Second, we adopt a similar transfer learning approach to Toonify [Pinkney and Adler 2020]. We sample latent codes in Z^+ from a multi-variate Gaussian distribution and then fine-tune an *attribute-aware generator* starting from StyleGAN2's pre-trained weights. The training losses include an adversarial loss with the given style exemplars, a facial structural loss [Zhang et al. 2018], as well as R1 [Mescheder et al. 2018] and perceptual path-length [Karras et al. 2020b] regularization losses. The attribute-aware generator includes multiple generative paths for different attributes (e.g. gender) and multiple discriminators to better capture attribute-dependent stylization of facial features. To avoid overfitting caused by a small training dataset, and to better balance identity and style, we adopt an early stopping strategy in training. During inference, the hVAE encoder and attribute-aware generator can be pipelined to generate stylized output from input.

Our approach provides greater agility in creating high quality and high resolution (1024×1024) portrait stylization models, requiring only a limited number of style exemplars (around 100) and short training time (around 1 hour). To evaluate our method, we collected

a number of artistically styled datasets, including 3D cartoons, oil paintings, comics and a few celebrity photos. We show that we can achieve superior quality to previous state-of-the-art portrait stylization methods, including Toonify [Pinkney and Adler 2020], with comparisons done qualitatively, quantitatively and through a perceptual user study. In addition, we demonstrate the ability to change 3D viewpoints and illumination using [Shen and Zhou 2020] in the latent space and to perform motion retargeting using [Siarohin et al. 2019] on the resulting stylized portraits. Finally, we include an ablation study on the key components of our system and demonstrate that our hVAE approach outperforms several state-of-art StyleGAN inversion methods in terms of inversion consistency.

To summarize, the key contributions of this paper are:

- A novel method that can agilely create high quality portrait stylization models with limited numbers (around 100) of unpaired style exemplars and short training time;
- An inversion-consistent transfer learning framework that solved the inversion discrepancy problem in transfer-learning-based stylization;
- A hierarchical Variational Autoencoder and its associated augmented Z^+ latent space that captures different levels of facial features for improved portrait stylization.

2 RELATED WORK

We review some existing researches that are relevant to the portrait stylization problem addressed in this paper.

Face Stylization. Stylizing facial images in an artistic manner has been explored in the context of non-photorealistic rendering. Early approaches relied on low level histogram matching using linear filters [Heeger and Bergen 1995]. Neural style transfer [Gatys et al. 2016], by matching feature statistics in convolutional layers, led to early exciting results via deep learning. Since then, several improvements have been proposed. Li et al. [2016] enforced local patterns in deep feature space via a Markov random field (MRF). Runder et al. [2016] extended style transfer to video and improved the quality by imposing temporal constraints. Although those methods can achieve generally compelling results for several artistic styles, they usually fail on styles involving significant geometric deformation of facial features, such as cartoonization.

For more general stylization, image-to-image (I2I) translation may be used to translate an input image from a source domain to a target domain. The seminal work here is "pix2pix" [Isola et al. 2017], which used a conditional generative adversarial network [Goodfellow et al. 2014] to learn the input-to-output mapping. Similar ideas have been applied to various tasks, such as sketches-to-photographs [Sangkloy et al. 2017] and attribute-to-image [Karacan et al. 2016]. However, these methods require paired training data, which is hard to obtain. To avoid this, conditional image generation may be approached in an unsupervised manner. For example, the well-known cycle-consistency loss in CycleGAN [Zhu et al. 2017] was proposed to improve network training stability for the unpaired setting. Unsupervised methods have also been used in cartoonization. Li et al. [2018] extended CycleGAN [Zhu et al. 2017] to cross-domain anime portrait generation. Kim et al. [2020] incorporated an attention module and a learnable normalization function for cartoon face generation, where their attention-guided model can flexibly control the amount of change in shape and texture. Although these methods can conduct plausible image translation, such networks require extensive training data, and thus most were only trained for relatively low image resolutions. Very recently, [Pinkney and Adler 2020] proposed a GAN interpolation framework for controllable cross-domain image synthesis, called Toonify, which can generate photo-realistic cartoonization. However, their inversion mapping when applied to real images often introduces undesired artifacts in the stylized output. In contrast, our proposed VAE inversion enhances distribution consistency in latent space, which leads to better results for real input images.

Generative Adversarial Networks (GANs). GANs have been used to synthesize images that ideally match the training dataset distribution via adversarial training. Starting from Goodfellow et al. [2014], GANs have been applied to various areas, e.g. image inpainting [Yuan et al. 2019], image manipulation [Bau et al. 2019a] and texture synthesis [Gecer et al. 2020]. Various advancements have been made to improve the architecture [Gulrajani et al. 2017], synthesis quality [Mao et al. 2017], and training stability [Mao et al. 2017]. However, initial methods only worked in low resolutions, due to computational cost and shortage of high-quality training data. Subsequently, a high-quality human face dataset, CelebA [Liu et al. 2015], was collected, and Karras et al. [2016] proposed ProGAN to train GANs for high resolution image generation via a progressive strategy; this can generate realistic human faces at a high resolution of 1024×1024. Recently, Karras et al. [2019] also collected a high resolution human face dataset called FFHQ, and, inspired by adaptive normalization for style transfer [Huang and Belongie 2017], proposed a new generator architecture StyleGAN to further improve face synthesis quality to the level that is almost indistinguishable from real photographs. Very recently, they [Karras et al. 2020b] extended this to StyleGAN2, which has reduced artifacts and improved disentanglement by using perceptual path length. Our work is built upon on StyleGAN2 and leverages their pre-trained weights as initialization.

GAN Inversion. Since GANs are typically designed to generate realistic images by sampling from a known distribution in latent space, GAN inversion addresses the complementary problem of finding the most accurate latent code, when given an input image, that will reconstruct that image. One approach is based on optimization [Abdal et al. 2019a; Karras et al. 2019; Tewari et al. 2020], directly optimizing the latent code to minimize the pixel-wise reconstruction loss for a single input instance. Another approach is learning-based [Zhu et al. 2016], in which a deterministic model is trained by minimizing the difference between the input and synthesized images. There are also some works that combine these ideas, e.g. learning an encoder that produces a good initialization for subsequent optimization [Bau et al. 2019b]. In addition to image reconstruction, some methods also use inversion when undertaking image manipulation. For example, Zhu et al. [2020] introduced a hybrid method to encode images into a semantic manipulable domain for image editing. Recently, Richardson et al. [2020] presented the generic Pixel2Style2Pixel (PSP) encoder. This is based on a dedicated identity loss for embedding images in several real image translation tasks, such as inpainting and super resolution. However,

these methods for single domain manipulation or reconstruction may not be directly applicable to cross-domain generation, due to insufficient consistency in the latent distributions, which is the issue addressed in our method. Our motivation is also related to PULSE [Menon et al. 2020], which involves GAN latent space exploration for photo upsampling. Here regularization is done by sampling from and constraining the exploration to a hypersphere prior in latent space. While this can improve the quality of super-resolved results, it does not prevent the sampling to be skewed from the underlying image distribution. In contrast, our approach is directly aligned with the variationally-derived principles of the VAE, and thus will lead to samples that collectively fit the ground truth image distribution, as mapped from the StyleGAN Gaussian latent Z prior.

Learning with Few Samples. Training a modern high-quality, high-resolution GAN typically requires 10^5 images, which is a costly undertaking in terms of acquisition, processing, and distribution. There are a few techniques to reduce such requirements. For example, Liu et al. [2019] and Wang et al. [2019] introduced a few-shot learning technique to perform appearance translation without needing a large dataset of specific style translation pairs. However, a pre-trained style embedding network is required and the generated image resolution is limited. Conversely, the idea of patch-based training [Shaham et al. 2019; Shocher et al. 2018] was further explored, as less training data is needed when learning patch distributions. However, such techniques may not easily be relevant to portrait generation, since human faces have strong geometry semantics and may not simply be reduced to smaller patches for training. To address the data shortage, our method is based on applying transfer learning to StyleGAN2, adopting an early stopping strategy to generate optimal results.

3 METHOD

Given a small set of stylistic exemplars, our goal is to design a pipeline that generates a high quality stylized images from real portrait images as input. The output image should be recognizable as the input subject's identity. It should also preserves subject's pose and expression. Last, it should be rendered in a style that is consistent with the provided stylistic exemplars.

The overall pipeline is shown in Fig. 3. The starting basis for our pipeline is a pre-trained StyleGAN2, and recall that if we took random samples from a Gaussian distribution in the Z latent space, it will generate images fitting the original training distribution (FFHQ).

In general, there are two major stages involved in our training pipeline. (A) Since our task involves using an image as input, we want to determine its corresponding latent vector for StyleGAN2. We therefore train a front-end encoder in a VAE setting to map input images into latent space, while keeping the back-end StyleGAN2 generator fixed. (B) Starting from a copy of the pre-trained StyleGAN2, we fine-tune this generator such that if we sample from a Gaussian distribution in latent space, it will generate images that better fit the stylistic exemplars given.

Notice that the two training stages are execution independent and can be trained in parallel. However, structurally the two stages have shared pivot latent spaces (Z+ and W+ described later), and are also jointly anchored by the fixed StyleGAN2 generator. There are three unique benefits of breaking down the entire generation problem into two stages: 1) the training does not require paired datasets, unlike typical image-to-image translation methods[Isola et al. 2017]. 2) the separation of training also enables higher resolutions by reducing computational load in making backpropagation more effective and efficient. 3) the compartmentalization of the pipeline allows greater agility, whereby new style domains can be incorporated by only fine-tuning the generator instead of the entire pipeline.

Over the following sections, we will introduce a new embedding space Z^+, and a hierarchical variational autoencoder (hVAE) using pyramid feature extraction to better embed and reconstruct human faces. We then present an attribute-aware generator with a multi-path structure, trained for synthesizing stylized images through transfer learning using a combination of GAN loss, similarity loss and regularization loss. Finally at inference stage, the encoder and stylized generator are combined to form a single-pass pipeline.

3.1 Z+ Space

The pre-trained StyleGAN model [Karras et al. 2019, 2020b] is equipped with two latent spaces: the original latent space Z under a Gaussian distribution, and a less entangled W space, which is mapped from Z through a Multi-Layer Perceptron (MLP) f. The original StyleGAN2 generation is conducted in a coarse-to-fine manner using several disentangled layers but with the same latent code input to each layer. Inspired by Image2StyleGAN [Abdal et al. 2019a] that enlarges StyleGAN's W space to $W+$ space, we chose to increase our model's expressiveness by using a different latent code from Z for each layer, allowing for individual control. This is equivalent to stacking multiple versions of the original latent space Z to form a new space Z^+. Unlike Image2StyleGAN [Abdal et al. 2019a] that targets pixel-level reconstruction by embedding into $W+$ space, we adopted Z^+ for two reasons. 1) our task involves cross-domain image generation. This makes it harder to directly embed into the $W+$ space without suffering deterioration in stylization quality, since we cannot assume all the codes in $W+$ are appropriate for stylization. 2) the $W+$ space is covered by a complex non-Gaussian distribution [Wulff and Torralba 2020] and directly encoding images into $W+$ via a network may not correspond appropriately to a Gaussian distribution in Z^+. Conversely, our stylization task is best addressed via Z^+ space, as the more constrained Gaussian modeling here leads to better regularization across different styles.

3.2 Hierarchical Variational Encoder

Hybrid Variational Autoencoder. In order to inverse map an input image back into the Z^+ latent space, we adopt a variational autoencoder (VAE) formulation [Kingma and Welling 2014; Rezende et al. 2014]. A typical VAE consists of an encoder $\mathcal{E}_\theta$ and a decoder $\mathcal{G}_\phi$ with respective parameters θ and ϕ, which are trained jointly to minimize reconstruction error between input image x and output image x'. Here, we instead propose a hybrid variational autoencoder for inversion that uses a *fixed* original pre-trained StyleGAN2 as the decoder $\mathcal{G}_{\phi_o}$, and we focus only on training the encoder network to learn the posterior distribution $q(z|x)$.

While a simpler alternative may have been to learn a deterministic encoder, a VAE approach improves robustness and generalization

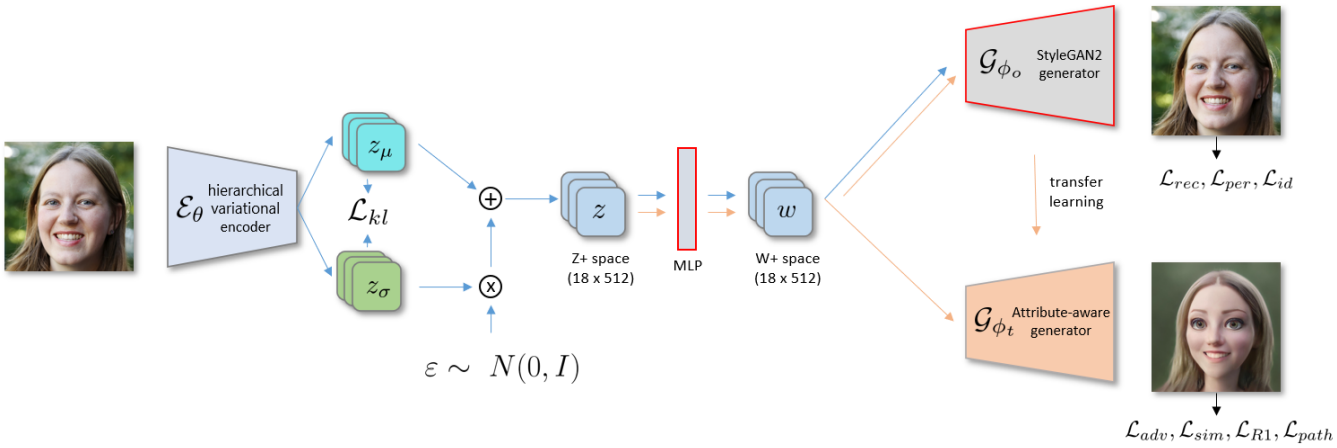

Fig. 3. Pipeline overview. Our hierarchical VAE consists of an encoder and generator with different color arrows representing the different training dataflows based on StyleGAN2. The blue arrows indicate image embedding, and the orange ones are for transfer learning. black borders indicate the block weights, which are derived from a StyleGAN2 pre-trained on the FFHQ dataset, that are frozen during training. The input is courtesy of Erin Wagner(Public Domain).

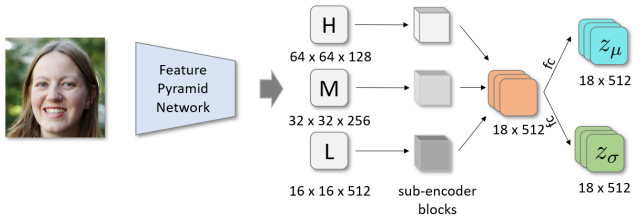

Fig. 4. Structure of our hierarchical variational encoder.

ability, as the Gaussian sampling component introduces useful perturbations during learning. We train the encoding parameters θ using the stochastic gradient variational Bayes (SGVB) algorithm [Kingma and Welling 2014] to solve:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{z\sim\mathcal{E}_\theta(x)} [-\log\ p(x|z)] + D_{kl}(\mathcal{E}_\theta(x)||p(z)), \quad (1)$$

where D_{kl} denotes the Kullback-Leibler divergence. The posterior / importance distribution, mapped by the encoder from x, is modeled as a multivariate Gaussian distribution $q(z|x) = \mathcal{E}_\theta(x) = N(z_\mu, \text{diag}(z_\sigma^2))$, where $z_\sigma, z_\mu \in \mathbb{R}^{18\times512}$ are the multi-dimensional output of $\mathcal{E}_\theta(x)$, representing the mean and standard deviation respectively in a diagonal matrix form. The prior is $p(z) = N(0, I)$, as used in StyleGAN2, and thus the KL divergence can be expressed in the analytic form of

$$D_{kl}(\mathcal{E}_\theta(x)||N(0, I)) = \frac{1}{2}\sum_i (1 + 2\log\ z_{\sigma,i} - z_{\mu,i}^2 - z_{\sigma,i}^2), \quad (2)$$

where the summation applies across all dimensions of z_σ and z_μ. Backpropagation is made differentiable via the reparameterization trick [Kingma and Welling 2014], whereby z can be sampled according to:

$$z = z_\mu + \epsilon \otimes z_\sigma, \epsilon \sim N(0, I), \quad (3)$$

where $\otimes$ is an element-wise matrix multiplication operator.

Hierarchical Feature Extraction. One unique aspect of StyleGAN2 is that the intermediate style codes mapped from Z^+ are injected into different layers of the generator and can semantically control image generation. The style codes broadly fall into three groups: 1) style codes lying in lower layers control coarser attributes like facial shapes, 2) middle layer codes control more localized facial features, while 3) high layer codes correspond to fine details such as reflectance and texture. One straightforward way to embed an input image is to directly estimate the combined latent code 18x512 z in $Z+$ from a fully connected layer. However, it turns out to be difficult to effectively train such a network.

To address this issue, we followed recent efforts [Lin et al. 2017] and [Richardson et al. 2020] in utilizing the hierarchy of a pyramid network to capture three levels of detail from different layers. Specifically, the input image at 256×256 resolution is passed through a headless pyramid network to produce three levels of feature maps at different sizes, corresponding to coarse, medium and fine details. Each level's feature map goes through a separate sub-encoder block to produce a 6×512 code. Finally, the combined 18×512 code can be passed to the fully-connected layers to generate the means and standard deviations to represent the Gaussian importance distribution in Z^+.

Loss Functions. Multiple loss functions are used in training our encoder network $\mathcal{E}_\theta$. We first use L_2 loss for reconstruction as follows:

$$\mathcal{L}_{rec} = \mathcal{L}_2(x, \mathcal{G}_{\phi_o}(\mathcal{E}_\theta(x))) \quad (4)$$

This measures the pixel-level differences between input image x and generated output $\mathcal{G}_{\phi_o}(\mathcal{E}_\theta(x))$. In addition, we utilize the LPIPS loss [Zhang et al. 2018] to learn perceptual-level similarities:

$$\mathcal{L}_{per} = \mathcal{L}_{lpips}(x, \mathcal{G}_{\phi_o}(\mathcal{E}_\theta(x))) \quad (5)$$

To preserve identity, we also use a facial recognition loss:

$$\mathcal{L}_{id} = \mathcal{L}_{arc}(x, \mathcal{G}_{\phi_o}(\mathcal{E}_\theta(x))), \quad (6)$$

where $\mathcal{L}_{arc}$ is based on the cosine similarity between intermediate features extracted from a pre-trained ArcFace recognition network [Deng et al. 2019], comparing the intermediate features of the source and output images. The KL divergence loss is defined as:

$$\mathcal{L}_{kl} = D_{kl}(\mathcal{E}_\theta(x)||N(0,I)), \tag{7}$$

In combination, our total loss becomes

$$\mathcal{L} = \mathcal{L}_{rec} + w_{per}\mathcal{L}_{per} + w_{id}\mathcal{L}_{id} + w_{kl}\mathcal{L}_{kl} \tag{8}$$

where w_{id}, w_{per}, w_{kl} are relative weights for the reconstruction loss, perceptual loss, identity loss and KL divergence loss respectively.

Implementation Details. The hVAE parameters were trained on the CelebA-HQ dataset [Lee et al. 2020; Liu et al. 2015], which contains 28,000 high quality face images. To reduce network parameters and computation load, the input images were down-sampled to 256×256, with down-sampling also applied to reconstructed images for computing the losses. The pre-trained StyleGAN2 used weights from the config-f 1024×1024 FFHQ model [Karras et al. 2020b]. The training parameters were fixed to be $w_{per} = 0.8, w_{id} = 0.8$,and $w_{kl} = 5 \times 10^{-4}$. We minimized the objective function for 20 epochs using the Rectified Adam solver [Liu et al. 2020]. We used a batch size of 16 and learning rate of 1×10^{-3} on two 32GB Tesla V100 GPUs.

3.3 Attribute-Aware Generator

To generate stylized portraits, we train a generator using a relatively small collection of stylistic exemplars. The generator is based on StyleGAN2, but enhanced with a multi-path structure to better adapt to different features corresponding to known attributes, such as gender. The structure is shown in Fig. 5. To mitigate the small dataset problem and better preserve user identity, we adopt transfer learning and an early stopping strategy to train the generator.

Stylization. The training stability of StyleGAN2's architecture, and the availability of high-resolution pre-trained models, have made it possible to achieve high quality cross-domain generation using transfer learning. As artistic portraits share obvious perceptual correspondences to real portraits, our method uses a StyleGAN2 model, pre-trained on the high-resolution real portrait FFHQ dataset [Karras et al. 2019], as the initialization weights. The network is subsequently transfer-learned on the smaller stylized dataset. There are three key benefits of using StyleGAN2 for stylization: 1) fine tuning can significantly reduce training data and time needed for high quality generation, compared to training from scratch, 2) StyleGAN2's coarse-to-fine generation architecture can support various artistic styles, including geometric and appearance stylization, and 3) the transfer-learned generator $\mathcal{G}_{\phi_t}(z)$ which is derived from the original model $\mathcal{G}_{\phi_o}(z)$ can form a natural correspondence when given the same latent codes, even with different generator parameters of ϕ. Hence when given an input image x, the inverse mapped latent code z can first be obtained from the VAE encoder, and then passed to different stylized generators (trained on different stylized datasets). This results in different stylized images, i.e. $\{\mathcal{G}_{\phi_1}(\mathcal{E}_\theta(x)), \mathcal{G}_{\phi_2}(\mathcal{E}_\theta(x)), \mathcal{G}_{\phi_3}(\mathcal{E}_\theta(x))...\}$.

Multi-Path Structure. Typically, when artists design characters, they often emphasize attribute-dependent characteristics to enhance appearance. For example, facial features may be exaggerated differently to accentuate femininity or masculinity (Fig. 6). Those attribute-dependent characteristics usually involve different facial geometric ratios as well as different facial features. Directly using the existing single-path StyleGAN2 structure and a single discriminator may not be best at distinguishing these attribute-dependent characteristics, while training several single-path generators to cater to different attributes will increase time and memory. For efficiency, we embed a multi-path structure within the same generator $\mathcal{G}_{\phi_t} = \{\mathcal{G}^k_{\phi_t}\}, k \in \mathbb{A}$ corresponding to the different attributes $\mathbb{A}$, while using multiple discriminators $D = \{D_k\}$. Since lower layers of the network guide coarse-level features like facial shapes, while higher layers affect facial reflectance and textures, the multi-path structure is more appropriately embedded within the lower layers. Nonetheless, this structure can also be placed into the higher layers, in situations where it may be more appropriate.

Loss Functions. The full objective comprises four loss functions to fine-tune the generator $\mathcal{G}_\phi$. We first use an adversarial loss to match the distribution of the translated images to the target domain distribution:

$$\begin{aligned}\mathcal{L}_{adv} = \sum_{k\in\mathbb{A}} \big(&\mathbb{E}_{y_k}[min(0, -1 + D_k(y_k))] + \\ &\mathbb{E}_{z\sim N(0,I)}[min(0, -1 - D_k(\mathcal{G}^k_{\phi_t}(z)))]\big)\end{aligned} \tag{9}$$

where y_k are target style images, classified by attribute k. To preserve recognizable identity of the generated image, we introduce a similarity loss at perceptual level, given by a modified LPIPS loss [Zhang et al. 2018]. Specifically, we discard differences from the first 9 layers of the VGG16-based LPIPS, and use the remaining differences from higher level layers. This helps in capturing the facial structural similarity, while ignoring local appearance variation.

$$\mathcal{L}_{sim} = \sum_{k\in\mathbb{A}} \sum_{i=9}^{30} \Big(\mathcal{L}^i_{lpips}(\mathcal{G}^k_{\phi_t}(z), \mathcal{G}_{\phi_o}(z))\Big), \tag{10}$$

To help improve training stability and prevent formation of artifacts, regularization terms are employed. For discriminators, we use R1 regularization [Mescheder et al. 2018].

$$\mathcal{L}_{R1} = \frac{\gamma}{2} \sum_{k\in\mathbb{A}} \Big(\mathbb{E}_{y_k}[\|\nabla D_k(y_k)\|^2]\Big), \tag{11}$$

where $\gamma = 10$ is the hyper-parameter for the gradient regularization. For the generator, we use a standard perceptual path-length regularization $\mathcal{L}_{path}$ [Karras et al. 2020b] from StyleGAN2, which aids reliability and behavior consistency in generative models.

Finally, the generator and discriminators are jointly trained to optimize the combined objective of:

$$\min_\phi \max_D \mathcal{L}_{adv} + w_{sim}\mathcal{L}_{sim} + w_{R1}\mathcal{L}_{R1} + w_{path}\mathcal{L}_{path}, \tag{12}$$

where $w_{sim} = 0.5, w_{R1} = 5, w_{path} = 2$ are relative weights for the adversarial loss, similarity loss, and regularization loss respectively.

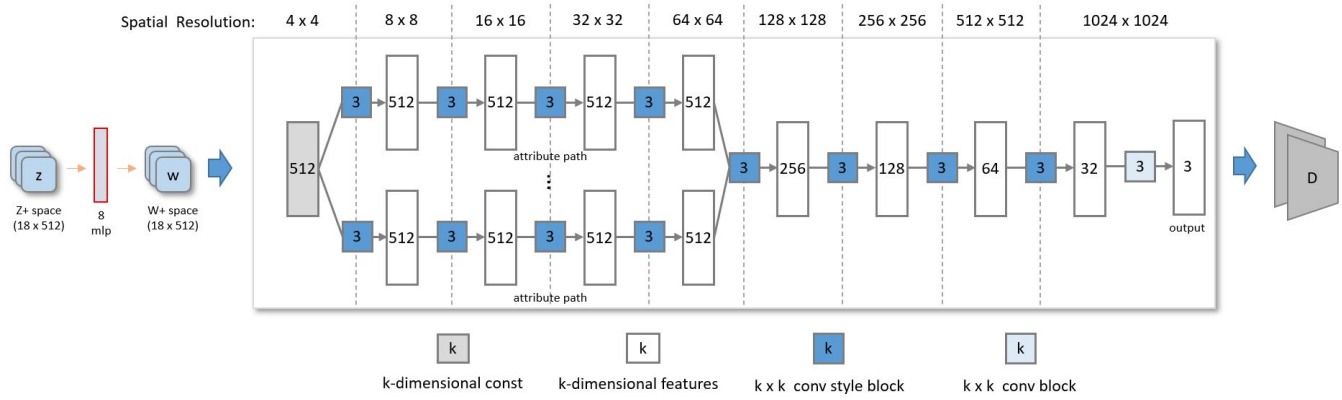

Fig. 5. The architecture of our attribute-aware generator network. Each latent code z, sampled from a standard Gaussian distribution, is first mapped to intermediate code w. Each w is forward into an affine transform in the style block [Karras et al. 2020b] and controls the generation via adaptive instance normalization (AdaIN) [Huang and Belongie 2017]. When decoding, a constant feature map is first initialized. Multiple paths are used in the lower layers for attribute specificity, while shared high layers unify texture appearance. Multiple attribute-specific discriminators are used to evaluate quality of the generated images. The network weights including discriminators are initialized from StyleGAN2.

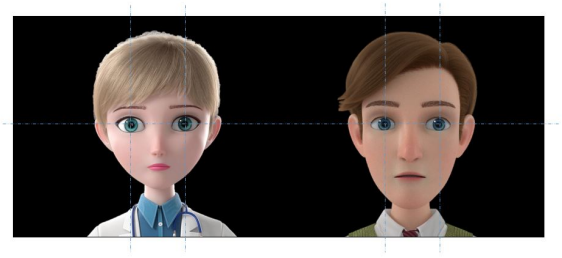

Fig. 6. Examples of attribute (gender) dependent accentuated differences in cartoons, including facial geometric ratios and features like eyelashes. The normalized exemplars are from our cartoon dataset.

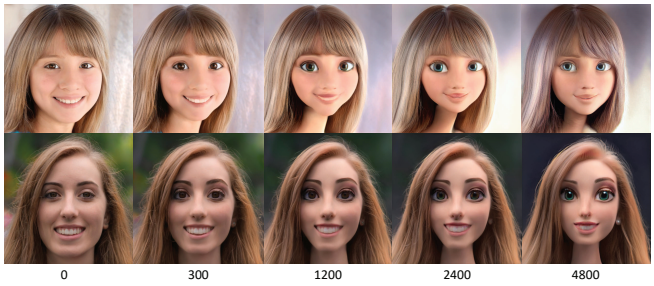

Fig. 7. Evolution of generative results after different training iterations. Input images are generated from StyleGAN2.

Dataset. We collected several artistic portrait images from multi-image asset websites, and also rendered images from 3D models. For the cartoon and comic styles, we rendered and collected some 100 images of each gender from two asset websites [pinterest 2021; turbosquid 2021]. For the oil painting style, 100 images of each gender were selected from the public Metfaces-dataset [Karras et al. 2020a]. For the two celebrity styles, we collected around 50 images from the internet, ignoring attributes. For each image, we extracted landmarks, conducted normalization by aligning eye positions, and cropped to a 1024×1024-sized image.

Early Stopping Strategy. A major potential problem with small datasets is that the discriminator may overfit the training examples, causing instability and degradation in GAN training [Karras et al. 2020a]. To mitigate this problem, we adopt an early stopping strategy to cease the training once the desired stylization effect has been achieved. As seen in Fig. 7, increasing the number of iterations further may also lead to increased deviation from the original input expression. To strike a balance between input fidelity and stylistic fit, we can cease training early, e.g. after 1200 iterations.

Implementation Details. The weights of the generator and discriminators are initialised based on the StyleGAN2 config-f 1024x1024 FFHQ model [Karras et al. 2020b]. This is then fine-tuned on a chosen style dataset, at a learning rate of 0.002 and with mirror augmentation. We used a batch size of 8 on two Tesla V100 32GB GPUs. The number of iterations usually ranges from 900 to 2200 due to the early stopping strategy, taking about 1 hour to train for each style.

3.4 Inference

Given an input face image x, it is first warped and normalized to 256×256 based on its landmarks, and encoded by the hVAE to get the latent Gaussian posterior distribution $q(z|x)$. Since this posterior / importance distribution is only relevant during hVAE training, we typically do not sample from this distribution during inference, but directly use the distribution mean as the latent code z, which will also better maintain temporal consistency. This z is then passed to the chosen stylized generator to generate a 1024×1024 stylized image. However, in rare cases there may be high frequency artifacts generated. In these cases, we can fall back on sampling multiple

Table 1. Preference scores and Fréchet Inception Distances (FID) for difference stylization methods. Preference scores are computed over 1K questions, while FID scores are computed from 2K images.

Algorithm	Preference Score ↑	FID ↓
Toonify[Pinkney and Adler 2020]	25.4	82.7
CycleGan[Zhu et al. 2017]	1.3	109.3
UNIT[Liu et al. 2017]	10.8	123.9
UGATIT[Kim et al. 2020]	4.6	104.7
Ours	**57.9**	**64.7**

instances from the imputed Gaussian distribution, leading to multiple output images. We can then select one without artifacts, either manually or choosing the one with the smallest average perceptual distance [Zhang et al. 2018] among the output images. For the gender attribute, a simple external pre-trained gender detector network [Eidinger et al. 2014] is used. In total, the inference stage takes around 130ms per image.

4 RESULTS

4.1 Artistic Portrait Generation

We present artistic stylization results from a variety of input images in Fig. 18 and Fig. 19. For each input image, we show several artistic styles. Results demonstrate that our method can robustly handle input images that represent a variety of skin tones, genders, face shapes and hair styles under different illumination conditions, correctly creating different stylization for those inputs. More results and test code are provided in the supplementary file.[1]

4.2 Comparisons

Qualitative Results. In Fig. 8, the results of our method can be compared to Toonify [Pinkney and Adler 2020] and other recent unpaired image translation techniques, including CycleGan [Zhu et al. 2017], UNIT [Liu et al. 2017] and UGATIT [Kim et al. 2020]. For Toonify, we used the authors' code and settings for training their transferred generator on our cartoon dataset. In their method, they used an optimization method [Karras et al. 2020b] for embedding an input image in latent space, and fed the corresponding code to the transferred generator. For the other three image translation methods [Kim et al. 2020; Liu et al. 2017; Zhu et al. 2017], we also used the respective authors' code and settings to train their networks on the CelebA-HQ training dataset and our cartoon dataset. Due to convergence difficulties and GPU memory limitations, those methods were not able to directly support 1024×1024 resolution, thus we kept their original sizes of 256×256 for training and up-sampled the output to 1024×1024 for comparison.

From Fig. 8, it can be seen that our method successfully cartoonized subjects with visually pleasing results. Toonify's results exhibit some visible artifacts such as unusual yellowish patches. As for the other unpaired image translation methods, besides not supporting higher resolutions, they also did not cope well when trained with limited exemplars.

[1] Accompanying material can be found at: https://github.com/GuoxianSong/AgileGAN.

Quantitative Results. We conducted a perceptual user study in which 100 participants were shown stylization results from different methods, and asked to select the best cartoonized images. Each participant was shown 10 questions randomly selected from a question pool containing 100 examples (using images with indices 0-99 in the CelebA-HQ dataset). Table. 1 shows that results from our proposed method had the majority preference.

Another metric to quantitatively evaluate generative quality is the Fréchet Inception Distance (FID) score [Heusel et al. 2017], which measures the visual similarity and distribution between two datasets of images. Each method generated stylized images from the CelebA-HQ dataset as input, and we computed the FID to the training cartoon dataset. We can see from Table 1 that our method also achieved the best performance on this metric, although it should be noted that since there are fewer than 5K images in the CelebA-HQ test set, FID scores may not be reliable.

Alternative Encoder Methods. We compared our hierarchical variational encoder to alternative embedding methods, including PSP encoder [Richardson et al. 2020], in-domain encoder [Zhu et al. 2020] and StyleGAN2 optimization [Karras et al. 2020b], by evaluating with these substitutes. For the PSP and in-domain encoders, we re-trained their model using the authors' original code and settings on the CelebA-HQ dataset. For the in-domain encoder, it also needed additional optimization to refine the latent code by minimizing pixel differences, perceptual loss and discriminator loss. We also compared to the iterative optimization proposed in StyleGAN2 [Karras et al. 2020b] and used in Toonify, as well as the enhanced iterative optimization of Image2StyleGAN [Abdal et al. 2019b] that uses a combination of pixel-level L2 and weighted perceptual losses.

Test images from the CelebA-HQ dataset were encoded by these alternative methods directly into the W^+ latent space as per their design, and fed to our transfer-learned generator. Please, note that transfer-learned generator is trained independently and not involved our hVAE encoder. In Fig. 9, it can be seen that our method, which encodes into the Z^+ latent space, created stylized images that are perceptually more pleasant. The structure and losses of the alternative encoders are geared towards image reconstruction rather than cross-domain consistency, which may lead to artifacts such as unnatural color patches and blur in texture. Although the in-domain encoder used a pixel-level discriminator to regularize the domain embedding for semantic manipulation, it is still unable to synthesize pleasantly styled images for a cross-domain generation task. We also computed the FID scores from all test images of CelebA-HQ to quantitatively evaluate the generation quality after each embedding method. From Table 2, it can be seen that our method achieved the best performance, although as mentioned previously, FID scores may not be reliable for only 2K images. Nonetheless, our hVAE can generally perform better than other embedding methods visually, reducing artifacts and improving stylized image quality. In the supplementary material, we provide further visual comparisons. In addition, we evaluate our encoder using a pre-trained Toonify model as a generator from the official repo of Pixel2Style2Pixel[Richardson et al. 2020] in Fig. 10. It shows our hVAE can improve generative results and contain less artifacts.

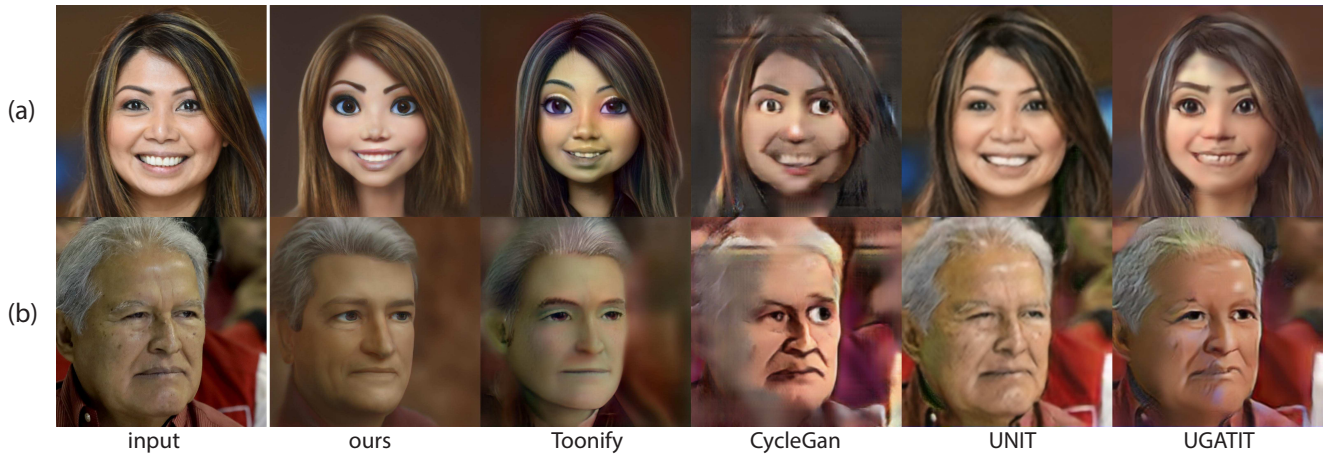

Fig. 8. Qualitative comparison of our method to state-of-the-art unpaired Image-to-Image translation techniques for cartoon style generation: Toonify [Pinkney and Adler 2020],CycleGan [Zhu et al. 2017], UNIT [Liu et al. 2017] and UGATIT [Kim et al. 2020]. Input images are courtesy of PFA SEAL(Public Domain) and Presidencia El Salvador(Public Domain).

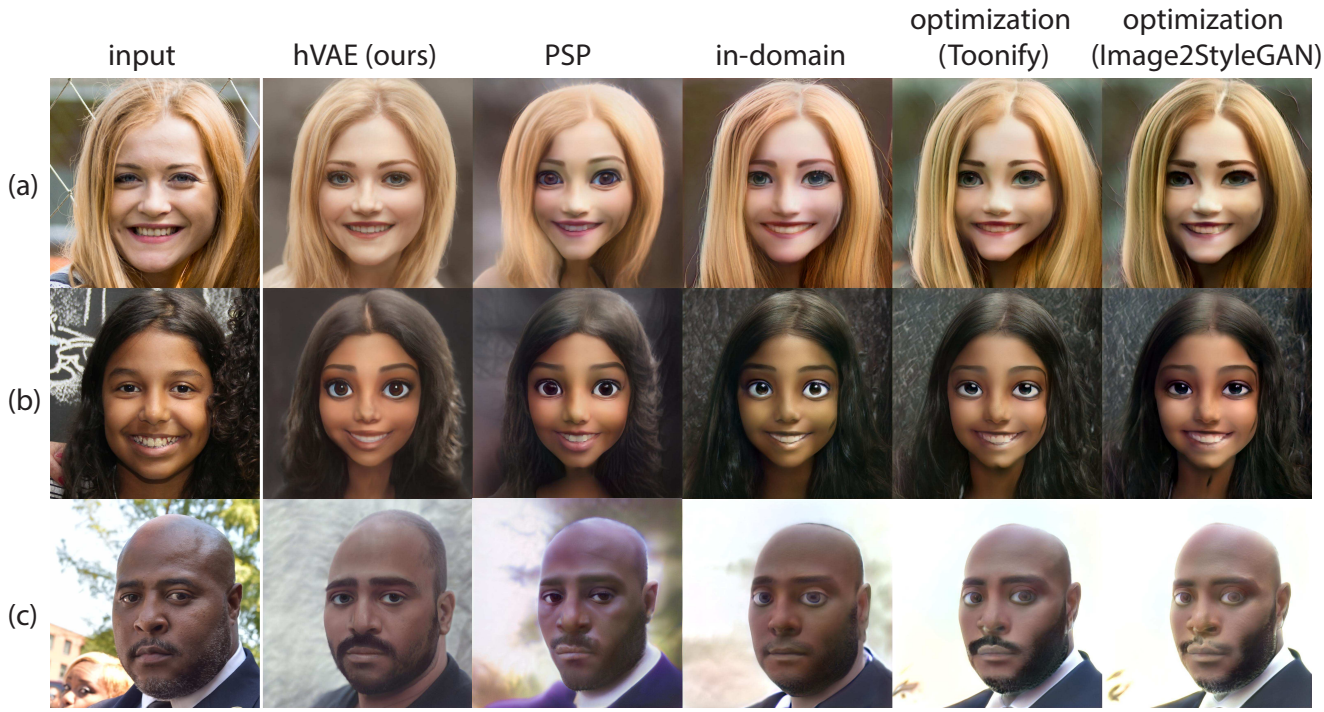

Fig. 9. Stylization results of our transfer-learned model when using the different encoder methods of PSP [Richardson et al. 2020], in-domain [Zhu et al. 2020], Toonify optimization (as proposed in [Karras et al. 2020b]) and Image2StyleGAN optimization [Abdal et al. 2019b]. Notice that results from other methods had greenish patches, blurry skin texture, and unnatural synthesis of eyes. Input images are courtesy of Ritvars Stankevičs(Public Domain),Alan kardek Ribeiro(Public Domain), and Daryl Levine(Public Domain). Also, there are more comparisons in supplementary.

Table 2. Fréchet Inception Distances (FID) for different embedding methods, computed from 2K images. Lower scores are better.

hVAE (ours)	PSP	in-domain	optimization (Toonify)	optimization (Image2StyleGAN)
64.7	69.0	78.6	71.9	74.7

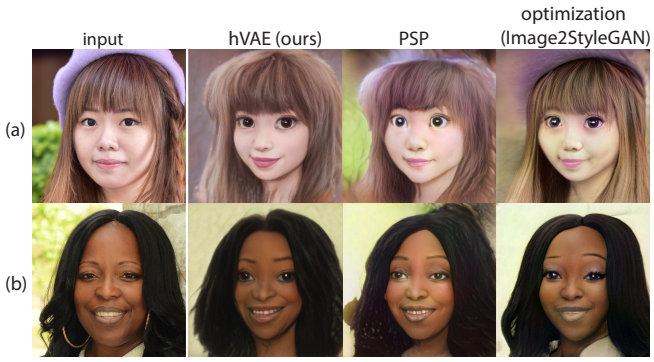

Fig. 10. Our hVAE encoder is universal, and can improve other stylization generator. Here, we compare hVAE with PSP [Richardson et al. 2020] and Image2StyleGAN optimization [Abdal et al. 2019b] using a pre-trained Toonify model as a generator from the official repo of Pixel2Style2Pixel[Richardson et al. 2020]. Input images are courtesy of barbybennett(Public Domain) and jang ba(Public Domain). More comparisons are presented in supplementary.

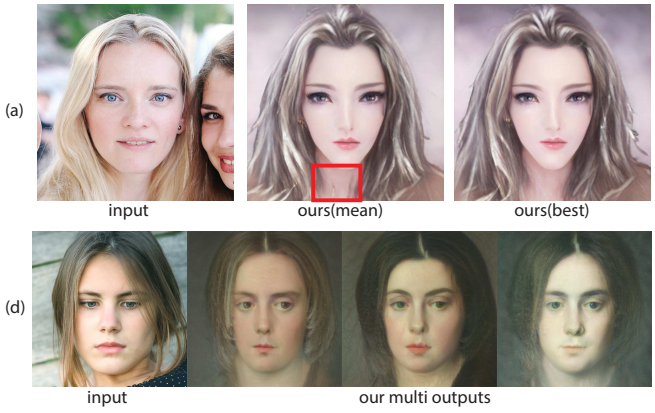

Fig. 11. Our pipeline allows for multiple sampling, allowing selection that avoids high frequency artifacts (a), and also to produce multiple diverse renderings of the same subject (c). Input images are courtesy of Pavel Savin(Public Domain) and svklimkin(Public Domain).

To better appraise the embedded latent distributions, we also visualized each embedded latent space with the t-distributed stochastic neighbor embedding (t-SNE) approach, shown in Fig. 2. For fair comparison, we visualized the distribution in the W^+ space. To get the original StyleGAN2 distribution, We first sampled the standard Gaussian distribution in Z^+ space 2000 times, and then mapped the samples through to W^+. For our method, we embedded CelebA-HQ test dataset using our encoder into Z^+, which were then mapped to W^+. For the other methods, the embedding of the test images were done directly in W^+ space. In Fig. 2, it can be clearly seen that our embedding distribution shares the greatest overlap with original one, which demonstrate the greater consistency of our method.

4.3 Ablation Studies and Further Investigations

Multiple Sampled Outputs. Compared to other methods, one advantage of our VAE-based approach is that we can generate multiple

Fig. 12. Qualitative results where different components in the encoder are ablated (without Z^+ augmentation, without variational encoding and our full method).Input images are courtesy of Great Place to Work Deutschland(Public Domain) and Colegio Vimagio(Public Domain). More comparisons are presented in supplementary.

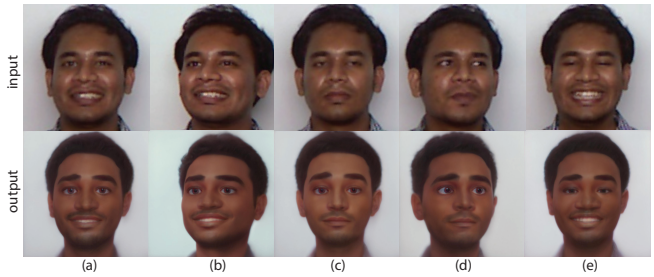

Fig. 13. Assessment of the quality and consistency of our method with different poses and expressions. Input images are from a multi-view expression dataset [Song et al. 2020].

results of the same person in a principled manner (without artificially adding latent noise) by sampling from the estimated posterior / importance distribution in Z^+. Fig. 11 showcases examples of multiple sampled results. In (a), artifacts may be present when coding with the mean of the posterior distribution, but better samples can easily be chosen manually. In (b), sampling can produce highly diverse but realistic results.

Hierarchical Variational Encoder. To further verify the usefulness of the designed modules in our hierarchical variational encoder, we conducted ablation studies by removing each component, with results shown in Fig. 12: (1) 'w/o Z+': replace the latent space Z^+ with Z. (2) 'w/o variational': remove the regularizer and replace variational encoding with a deterministic encoder. For each study, we retrained the encoder using the same settings, and also used same generator. For 'w/o Z+', embedding into Z space led to insufficient reconstruction ability and expressiveness. For 'w/o variational', we can see using a variational encoder can generate perceptually better stylization results with reduced artifacts, even for challenging input images with occlusion.

Different Poses and Expressions. We also qualitatively assess the generative quality and consistency of the stylized identity across different facial poses and expressions. We generated several results

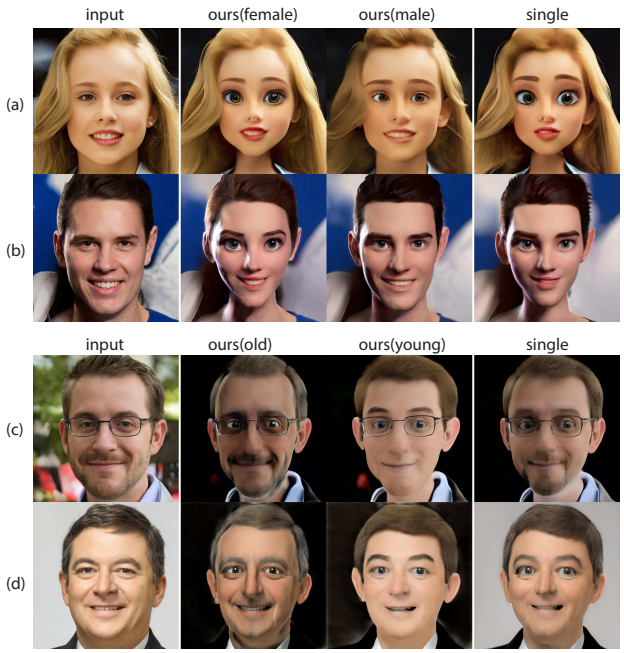

Fig. 14. Stylization results for our multi-path attribute-aware model compared to a single-path model. (a, b) are related to the gender attribute using multi-path at low layers, while (c, d) are related to age using multi-path at high layers. Please note attribute-sensitive features such as eye lashes and wrinkles. Input images are generated from StyleGAN2.

Fig. 15. Stylization results compared to Toonify in two different settings: fine-tuned only and layer swapping. Input images is courtesy of U.S. Department of Agriculture(Public Domain).Please zoom in to see difference.

of the same individual with different facial poses and expressions, as presented in Fig. 13. Our method performed well with poses and expressions retained from the input images, and the stylized character is recognizably the same across these variations.

Multi-Path Structure. To evaluate the usefulness of our multi-path generation module, we replaced it with a single-path structure and retrained the network on our cartoon dataset. From Fig. 14, we can see that using a dual-path structure can better generate gender-associated facial features in terms of facial geometry and length of eye lashes, and also enhance age-based features such as wrinkles.

Fine-tuned Only vs Layer Swapping. Instead of using a fine-tuned generator directly, Toonify's stylization is done by additionally swapping or blending in higher layer weights from the original StyleGAN2. We also investigated this approach for our framework, which trades off stylization for increased realism. Fig. 15 compares our

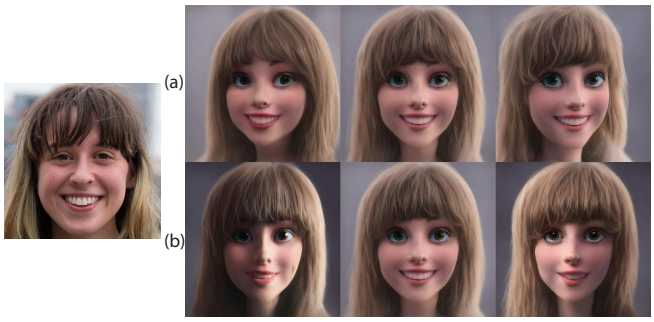

Fig. 16. Our pipeline also supports image editing, with changes in: (a) pose, and (b) illumination direction. Input image is courtesy of Mark Dixon(Public Domain).

method to Toonify under two settings: when only the fine-tuned stylized model is used, and when layer swapping is used. We can see that our method produced perceptually better stylization results and contained less artifacts.

4.4 Applications

Image Editing. Our pipeline can also support image editing via latent code manipulation, e.g. for semantic editing. Fig. 16 presents samples of pose and illumination editing. We used a closed-form method [Shen and Zhou 2020] to extract semantic directions (e.g. corresponding to pose and illumination changes) in W^+ from our stylized generator. Results presented are based on extrapolation in the semantic directions in W^+. Further results can be seen in the supplementary video.

Video Results. Our results can also be converted into video sequences via the first order motion technique [Siarohin et al. 2019], with the input video driving the stylized image. Please note that the pre-trained first order motion model only supports 256×256 frames. We refer readers to the supplementary video.

5 CONCLUSION

In this paper, we presented AgileGAN, a framework that can generate high quality stylistic portraits. Our method can agilely create high resolution portrait stylization models with a limited number (around 100) of unpaired style exemplars and with a short training time. This is done through a novel inversion-consistent transfer learning framework that reduces the issue of inversion distribution discrepancy. We also introduce a hierarchical Variational Autoencoder and its associated augmented Z^+ latent space that captures different levels of facial features for improved portrait stylization.

Limitations. Even though we presented a large variety of compelling portrait stylization results, there is still room for further improvement in our approach. Fig. 17 shows some examples. (a) Our method may fail to preserve accessories such as earrings or glasses after translation, since such cases are under-represented in the style datasets. (b) In some cases, we found that the generated gaze direction may not be consistent with the input, as training images are generally biased towards frontal gaze. (c) Classical oil paintings tend to have neutral face expressions, so likewise the more intense

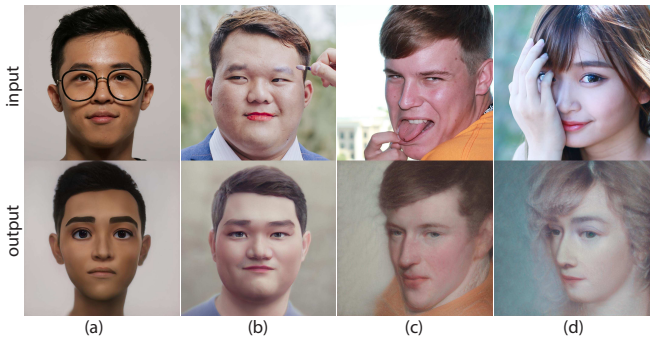

Fig. 17. Examples of failure cases. (a) missing accessories (glasses), (b) different gaze directions, (c) different expressions (for oil paintings), (d) heavy occlusion. Images are courtesy of BILL LIU(Public Domain), Luna(Public Domain), ijohn9n9(Public Domain) and RONG(Public Domain).

expressions may not be reproduced. We believe these problems can be mitigated by using more diverse datasets, where possible. (d) For extremely abstract styles or input with heavy occlusion, our method may not be able to to generate sufficiently accurate results.

Future Work. There are many future avenues to extend our work. (1) While cross-domain editing has been demonstrated in Fig. 16 and supplementary video, the latent space interpolation does not lead to smooth variation in the hair. One possible research is to improve the editing for consistency across such interpolated frames. (2) Another natural step is to extend the current single-image generation to video. We demonstrated some results using the first-order motion technique, but only low resolution videos can be generated for now, due to the use of pre-trained first-order motion models. It may be interesting to investigate how motion consistency can be more fundamentally integrated into our current pipeline for high resolution styled video generation.

ACKNOWLEDGMENTS

We would like to thank all the anonymous reviewers for their insightful discussion and valuable suggestions to improve quality of this paper. We also thank you Zhili Chen and Xiao Yang for discussion and Matthew Goodman for voice-over.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019a. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. In *ICCV*.

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019b. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. In *ICCV*.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, JunYan Zhu, and Antonio Torralba. 2019a. Semantic Photo Manipulation with a Generative Image Prior. In *ACM Transactions on Graphics*.

David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. 2019b. Seeing What a GAN Cannot Generate. In *ICCV*.

Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*.

E. Eidinger, R. Enbar, and T. Hassner. 2014. Age and Gender Estimation of Unfiltered Faces. *IEEE Transactions on Information Forensics and Security*.

L. A. Gatys, A. S. Ecker, and M. Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *CVPR*.

Baris Gecer, Alexander Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. 2020. Synthesizing Coupled 3D Face Modalities by Trunk-Branch Generative Adversarial Networks. In *ECCV*.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proc. NeurIPS*.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved Training of Wasserstein GANs. In *NeurIPS*.

David J. Heeger and James R. Bergen. 1995. Pyramid-Based Texture Analysis/Synthesis. In *ACM Trans. Graph.*

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.).

Xun Huang and Serge Belongie. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In *ICCV*.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-image Translation. In *ECCV*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.

Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. 2016. Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts. In *Proc. NeurIPS*.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020a. Training Generative Adversarial Networks with Limited Data. In *Proc. NeurIPS*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020b. Analyzing and Improving the Image Quality of StyleGAN. In *CVPR*.

Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. 2020. U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. In *International Conference on Learning Representations*.

Diederik P. Kingma and M. Welling. 2014. Auto-Encoding Variational Bayes. (2014).

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *CVPR*.

Chuan Li and Michael Wand. 2016. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis. In *CVPR*.

Jerry Li. 2018. Twin-GAN – Unpaired Cross-Domain Image Translation with Weight-Sharing GANs.

T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017. Feature Pyramid Networks for Object Detection. In *CVPR*.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the Variance of the Adaptive Learning Rate and Beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised Image-to-Image Translation Networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. 2019. Few-shot Unsueprvised Image-to-Image Translation. In *CVPR*.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*.

X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. 2017. Least Squares Generative Adversarial Networks. In *ICCV*.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In *CVPR*.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. 2018. Which Training Methods for GANs do actually Converge?. In *International Conference on Machine Learning (ICML)*.

Justin N. M. Pinkney and Doron Adler. 2020. Resolution Dependent GAN Interpolation for Controllable Image Synthesis Between Domains. In *NeurIPS Workshop*.

pinterest 2021. pinterest. https://www.pinterest.com/.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference on International Conference on Machine Learning*.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2020. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. *arXiv preprint arXiv:2008.00951* (2020).

Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. 2016. Artistic Style Transfer for Videos. In *German Conference on Pattern Recognition*.

P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays. 2017. Scribbler: Controlling Deep Image Synthesis with Sketch and Color. In *CVPR*.

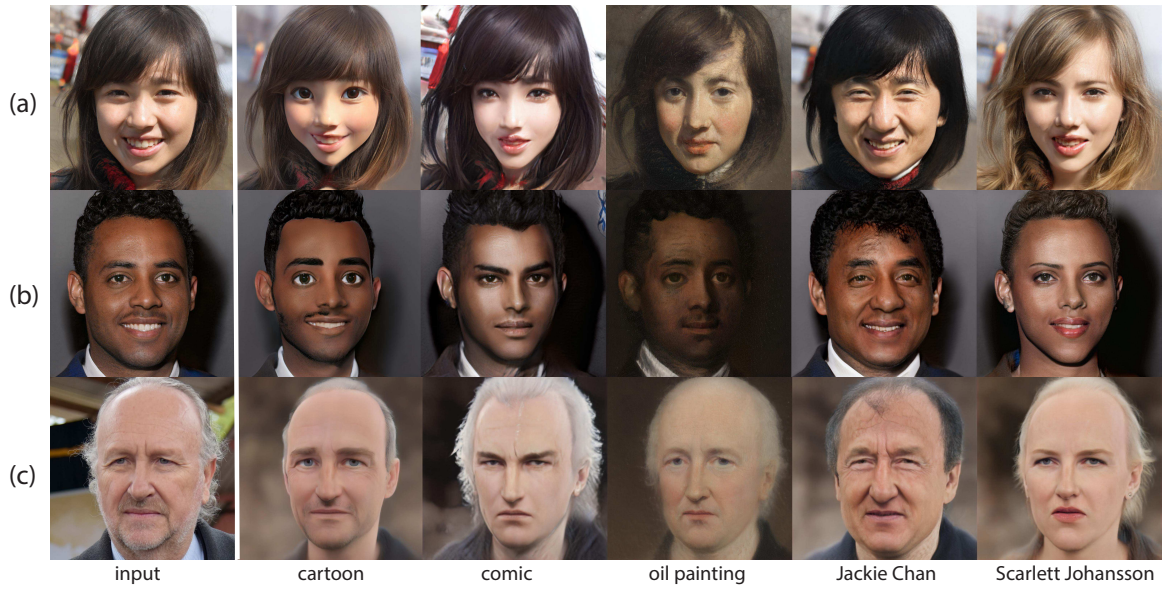

Fig. 18. Artistic portrait generation results from a variety of input images. From left to right, we show the input image, and generated stylistic images using our pipeline. For the input images (a)(b) are from StyleGAN2, while (c) is real image courtesy of Ministerio Minería(Public Domain). Please zoom in to see details.

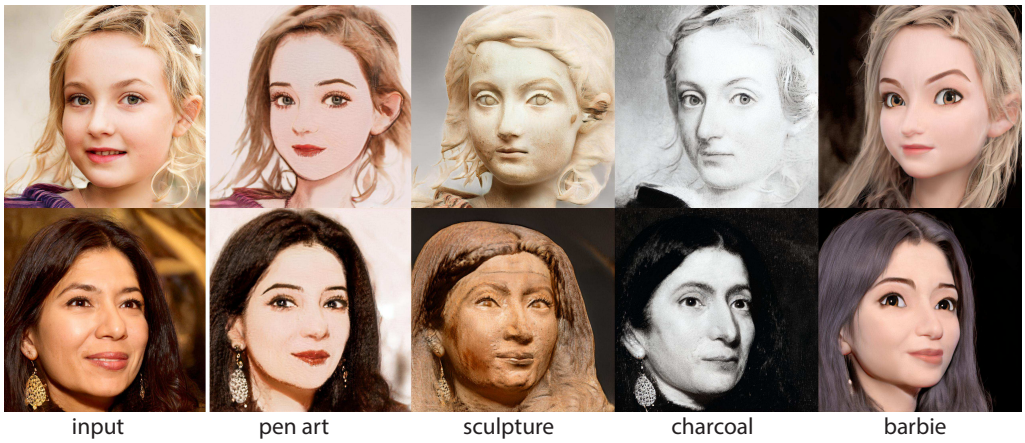

Fig. 19. Artistic portrait generation results from a variety of input images. Input images on the left are from StyleGAN2. The other columns are our generated results in four additional styles, with each style trained on either a category of MetFaces[Karras et al. 2020a] or a self-collected dataset. Please zoom in to see details, and more examples are in the supplementary material.

T. R. Shaham, T. Dekel, and T. Michaeli. 2019. SinGAN: Learning a Generative Model From a Single Natural Image. In *ICCV*.
Yujun Shen and Bolei Zhou. 2020. Closed-Form Factorization of Latent Semantics in GANs. In *ECCV*.
A. Shocher, N. Cohen, and M. Irani. 2018. Zero-Shot Super-Resolution Using Deep Internal Learning. In *CVPR*.
Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *NeurIPS*.
Guoxian Song, Jianmin Zheng, Jianfei Cai, and Tat-Jen Cham. 2020. Recovering facial reflectance and geometry from multi-view images. In *Image and Vision Computing*.
Ayush Tewari, Mohamed Elgharib, Mallikarjun B R, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020. PIE: Portrait Image Embedding for Semantic Control. In *ACM Trans. Graph.*
turbosquid 2021. turbosquid. https://www.turbosquid.com/Search/3D-Models/.
Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. 2019. Few-shot Video-to-Video Synthesis. In *NeurIPS*.
Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *CVPR*.
Jonas Wulff and Antonio Torralba. 2020. Improving Inversion and Generation Diversity in StyleGAN using a Gaussianized Latent Space. In *Conference on Neural Information Processing Systems*.
L. Yuan, C. Ruan, H. Hu, and D. Chen. 2019. Image Inpainting Based on Patch-GANs. In *IEEE Access*.
Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.
Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-domain GAN Inversion for Real Image Editing. In *ECCV*.
Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. 2016. Generative Visual Manipulation on the Natural Image Manifold. In *ECCV*.
Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *ICCV*.