HiStyle: Reinventing the Historic Portrait via 3D-aware GAN

Student Name: Rong Yang
Email:ryang24@pembrokehill.org
School: Pembroke Hill School
Address: Kansas City, United States
Advisor Name: Zhu Li
Orgnization: University of Missouri-Kansas City

Submission Date : 01/08/2023

HiStyle: Reinventing the Historic Portrait via 3D-aware GAN

Rong Yang
ryang24@pembrokehill.org

Abstract

Restoring and reinventing historical portraits has long been a challenging task in computer vision. In order to faithfully reinvent the portrait images, it is necessary to not only restore the color and reconstruct the 3D geometry, but also extend to diverse artistic styles. Existing methods for each specific task have made huge progress. However, they struggle with conflicts between low-quality and accurate restoration, which hinders their ability to meet all the criteria in a unified model. To achieve these goals, HiStyle is proposed by this project, a novel generative model for reinventing historic portraits, simultaneously supporting 2D to 3D reconstruction, gray to RGB conversion, and photo-to-stylized image transformation. To achieve this, HiStyle proposes a Generative Adversarial Network (GAN) inversion technique to transfer a gray historical portrait into the latent space of a 3D generator, restoring the lost color information and meanwhile lifting the 2D image to 3D representation. Besides, the powerful CLIP text-to-image large language model is incorporated into 3D-aware GANs to realize zero-shot text-driven style transformation. To support more diverse styles, we additionally explore the power of the latent diffusion model to synthesize multiple 2D style extensions of the colorized images. The results demonstrate significant improvements in the quality and diversity of the generated images compared to existing methods for each specific task. They also highlight the potential of 3D-aware GANs for preserving cultural heritage.

1. Introduction

Historical portraits represent an important aspect of cultural heritage, as they provide a glimpse into the past and serve as a visual record of our history. As the Metaverse opens a potential application to create immersive virtual experiences with the interaction with historical people of different periods, a novel multi-task of reinventing historical portraits is proposed by this project, required to fulfill the following aspects: 1) **colorization**, i.e., attaching color to old portraits can make it easier to visualize the past and connect with history while enhancing the emotional impact of historic portraits with visually relatable and realistic. 2) **stylization**, i.e., the historic portrait stylization is a form of artistic expression that can be extended to multiple media, allowing for creative interpretation and experimentation with different styles and color schemes. 3) **3D lifting**, i.e., 3D representations of historical images can help to gain a better understanding of historical people, and can also help the general public engage with history in a more meaningful way. However, the process of restoring and reinventing historical portraits can be challenging due to the limitation in quality, such as grayscale or low resolution.

With the rapid development of deep learning techniques, various deep-learning models have been proposed to effectively address each problem individually, ranging from early brute-force networks to more recent and meticulously designed GAN [21]. For colorization, current Convolution Neural Networks (CNN) based methods [11, 17, 31, 63] are capable of the generation of color images that are visually plausible and consistent with the original scene, while also allowing for some degree of artistic freedom. However, for the task of portrait, the general methods of colorization fail to consider the diversity of humans and tend to paint the portraits in a similar color, due to a lack of learning on the distribution of human faces. Stylization methods [20, 24, 26, 33, 34, 57] are able to render the image in the target domain while preserving the source content, but they face the dilemma between the natural appearance of styles and the desire for data. To balance the aforementioned two parts, several works [23, 27, 32] focus on the single style, while it limits to the incapacity of adapting to diverse styles in real-world applications. Current methods of 3D reconstruction can create digital 3D models of real-world objects with the help of multi-view input images, while it is a task of single-view reconstruction for historic portraits. Challenges of single-view reconstruction remain in terms of accurately capturing fine details, and handling complex geometries. The problem has been even more severe when handling incomplete or noisy input data, such as historical images with low quality.

As the pipelines of the three different tasks are usually diverse and hard to combine into a general model for historic portraits, it is natural to ask a question: can a model be built that can restore the color and 3D shape, and meanwhile empower us to generate the extension of portraits in diverse styles? In this project, a novel approach is presented to generally fulfill the **colorization**, **stylization**, and **3D representation** that leverages the power of a pretrained text-image model and the recent advances in 3D-aware image representation.

First, the method proposed by this project is a GAN inversion technique to transfer a gray historical portrait into the latent space of a 3D generator, allowing us to restore the color information that has been lost over time. Meanwhile, the inherent representation of the 3D-aware generator lifts the 2D image into 3D coordinates, reconstructing the geometric shape of the original portrait. Besides, the 3D generator is further fine-tuned using the guidance of the CLIP model [45] to transfer the style from a source domain, such as a realistic photograph, to a target domain, such as painting or fantasy. By combining the strengths of the text-image model and 3D-aware image representation, a technique has been developed that can transfer the original 2D historical image to a 3D colorful image. This is a significant advancement over existing methods for reinventing historic portraits, as it allows us to better preserve the cultural heritage of historical portraits by providing a more complete and diverse representation of portraits.

However, the pipeline still faces two challenges, **the requirement of diverse data** and **the diversity of generated images in the style domain**. To transfer the portrait into arbitrary styles, collecting diverse data on target styles is pretty hard. Although CLIP-based [2, 10, 19, 42] methods enable the 3D-aware model to realize zero-shot stylization under the text guidance, the deterministic embedding of the CLIP text encoder in the text prompt results in a significant loss of diversity. As a result, the generative models in the target domain may not retain the same level of sample diversity as those in the source domain. To address the above challenges in our pipeline, this project additionally proposes to augment the data on target styles for each specific portrait after colorization. The style images can enhance the style diversity but are hard to collect, while the text-driven methods release the data dependence but undermine the diversity of the generative model in target domains. Therefore, the text to synthesize multiple 2D style extensions of the colorized images is utilized via the power of latent diffusion model [48]. With the use of the augmented target images, the fine-tuning direction determined by the text is disturbed and thus guides the generated images to various targets for different specific identities. The process of 3D domain adaptation is then conducted while ensuring both diversity in the text and releasing the dependency on image data.

Thanks to the proposed specific data augmentation, our zero-shot 3D-aware generative model, dubbed HiStyle, overcomes the challenges of the loss of diversity and fulfills all three aforementioned aspects of a historic reinvention task. In summary, our main contributions are:

- A novel network architecture is proposed for multi-task portrait reinvention, which enables colorization, style transfer, and 3D lifting in a unified framework.
- The diffusion model to 3D-aware GANs is introduced for style data augmentation, solving the problem that the text-driven generation lacks diversity in the target style domains.
- The gray-inversion loss is proposed for the GAN inversion of gray images, better restoring the natural color of historic portraits.

2. Related Work

2.1. Image Colorization

The task of image colorization for historic images aims to automatically add color to gray-scale images. One approach to automatic colorization is learning-based model [11, 17, 31, 63]. Deep colorization is one of the first methods to introduce the deep neural network to the task of colorization. To better learn the global and local information of the input image, several works [25, 60, 64] proposed to use two-branch network architecture to individually handle the semantic maps and global content. However, the above colorization models struggle with the task of coloring images containing multiple objects. In order to address this problem, Su et al. [51] introduced the pre-trained detection model to realize an instance-aware image colorization approach. The disadvantage of this approach is that different color targets may have the same gray value, so the color output will be an average result and the diversity of color will be lost when these models process targets with multiple colors.

To restore the natural color, another approach is to learn the color prior to each specific class via adopting GAN. These methods [7, 12, 28, 36, 37] learn a generator to generate diverse colors of an image and inverse the input grayscale images into its latent space. The GAN inversion method has shown promising results in achieving high-quality colorization of images with complex textures and multiple objects. However, the GAN inversion is an unstable reconstruction that may fail to faithfully maintain the details.

2.2. Style Transfer

Style transfer can be classified into image-reference methods and text-guided methods in terms of the indication of target style. Early methods for the image-reference style transfer [20, 24, 26, 33, 34, 57] has been proposed based on

CNN. They are capable of synthesizing the image of the target style while keeping the source content. With the gradual development of GANs, plenty of methods aim at the style domain adaptation of pre-trained 2D generators with limited training images. However, these image-guided models are style-specific and constrained to the style which can easily collect image data. To release the requirement of image data in the style domain, text-guided domain adaptation methods have been proposed for generative models with the power of CLIP and diffusion model. StyleGAN-Nada [19] proposes a fine-tuning scheme with CLIP-directional loss that utilizes the text embedding to shift the domain of a pre-trained generator. To get rid of the time-consuming optimization phase of StyleGAN-Nada, HyperDomainNet [2] and HyperStyle3D [10] leverage a hyper-network to directly predict the parameter offset of generators. Despite the impressive results of style transfer and shape deformation, CLIP-based methods suffer from the limited diversity among different identities due to the determinant target direction guided by the textual loss. Furthermore, DATID-3D [29] introduces the text-to-image diffusion model to generate and filter the amount of style image data for the training of generators, both enhancing the diversity and releasing the requirement of manually collecting data. However, the results of DATID-3D severely depend on the filter of style images synthesized by the diffusion model.

2.3. Single-image 3D Face Reconstruction

3D Reconstruction of a face from a single 2D image is a long-standing challenge due to the insufficient information in the input 2D image. With the publication of 3D Morphable Models (3DMM) [3] the first general face representation, many following works [1, 4–6, 13, 43, 44] focus on the method that estimates the parameters of 3DMM for fitting the 3D model to the input 2D image. Furthermore, to alleviate the 3D data dependence, learning 3D prior from 2D data is a natural idea that followed by recent deep learning techniques [52, 53, 55, 56]. Besides the parameterized model, another approach to learning face prior from 2D images is the 3D-aware GAN [8, 9, 15, 18, 22, 30, 38–41, 46, 49, 50, 59, 61, 62, 65]. With the intrinsic NeRF-based 3D representation, these generative models can render high-quality face images in arbitrary viewpoints with multi-view consistency. Additionally, several recent works propose to leverage the power of the diffusion model to imagine the 3D model from a single image. In particular, Rodin [58] explored the combination of tri-plane representation and diffusion model, achieving to generate realistic and editable head avatars.

3. Method

Given a historic gray image $\mathcal{I}_{\text{gray}}$, we aim to colorize the image via 3D-aware GAN inversion and then stylize the colorized images under the guidance of text prompts. To this end, we design a two-stage framework that fine-tunes the 3D generator G_θ trained on a source domain, to a new target domain without the requirement of any 3D or 2D image training data. We first leverage a method for GAN inversion to transfer the input gray images $\mathcal{I}_{\text{gray}}$ into the W space of the generator and design a specific loss to restore the colorful image $\mathcal{I}_{\text{color}}$. Second, we adapt the pre-trained 3D generator across different style domains to generate the stylized images $\mathcal{I}_{\text{style}}$ under the guidance of the target style images. Finally, to release the need of target-style images, we employ a pre-trained text-to-image diffusion model to synthesize several style images of the original colorized images $\mathcal{I}_{\text{color}}$ which is used to enhance the diversity of style transfer. The details are described as follows. The overview of our architecture is shown in Fig. 1.

3.1. Preliminaries on EG3D

In our work, we build upon the concept of 3D-aware GANs that employ NeRF [35] as the underlying 3D representation. Several studies have explored this approach, including works such as Graf [49], Pi-GAN [9], GIRAFFE [39], and StyleNeRF [22]. For our implementation, we draw inspiration from EG3D [40], one of these 3D-aware models, and utilize its pre-trained generator as the source domain model.

The generator in EG3D adopts a tri-plane representation to convey the 3D scene information and utilizes an implicit function to decode the tri-plane feature into volume density and color in 3D coordinates. Specifically, this function takes the tri-plane feature sampled by a 3D coordinate $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$ as input and generates per-point volume density $\alpha(\mathbf{x}) \in \mathbb{R}^+$ and view-dependent color $c(\mathbf{x}, \boldsymbol{\xi}) \in \mathbb{R}^3$, where $\boldsymbol{\xi}$ represents the sampling camera pose. Instead of directly computing pixel colors for image generation, EG3D performs volume rendering along the corresponding camera ray to compute a low-resolution feature map $f(\mathbf{x}, \boldsymbol{\xi}) \in \mathbb{R}^N$. To synthesize high-resolution images while considering memory constraints, EG3D efficiently upsamples the low-resolution feature maps using a super-resolution network.

In our method, the image generation process can be summarized as $\mathcal{I} = G_\theta(w, \boldsymbol{\xi})$, where G_θ represents the generator, w denotes the latent code in the generator's W space. By leveraging the capabilities of EG3D and incorporating our proposed modifications, we enhance the generation of high-quality and diverse historical portraits.

3.2. Colorization

The goal of colorization is to inverse the gray image of a historical portrait to the latent space of 3D-aware GAN which can imagine the image color by the learning from the face dataset. As Fig. 1 shows, our colorization contains

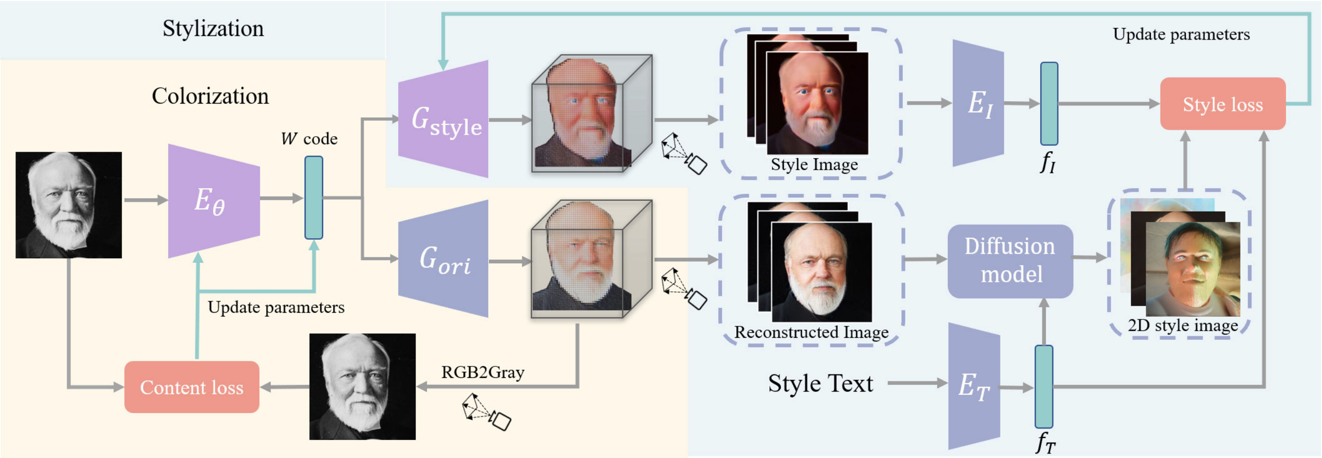

Figure 1. **The overview** of our full-pipeline. (a) Our colorization stage consists of an encoder to project the input image into W code and a fine-tuning procedure to further optimize the W code and the generator. (b) Our stylization stage leverages the latent diffusion model to construct a set of style images guiding the style transfer together with the target style prompt.

two parts, an encoder E to predict an initial w code and an optimization process to adapt the generator G_θ to the input image domain.

Encoder. Inspired by the previous GAN inversion for StyleGAN, our Histyle model utilizes the e4e encoder [54] as the backbone of E. Unlike the previous encoders for 2D GANs, the viewpoint $\boldsymbol{\xi}$ of the input image is needed as the additional input of the encoder to supply the 3D-aware knowledge for the generator. The inversion procedure can be described as,

$$w = E\left(\mathcal{I}_{\text{gray}}, \boldsymbol{\xi}\right), \tag{1}$$

where $\boldsymbol{\xi}$ can be obtained by an off-the-shelf 3D face reconstruction method [16]. However, the encoder tends to project the input gray image to a colorless portrait to ensure the reconstruction accuracy, but fails to reconstruct the color of generated images. To enable the encoder E to learn the gray image inversion for imagining the color, we additionally use an RGB-to-gray loss that encourages the reconstruction with color imagination,

$$L_{\text{gray}} = \left\| T\left(G_{\hat{\theta}}\left(w, \boldsymbol{\xi}_i\right)\right) - \mathcal{I}_{\text{gray}} \right\|_2, \tag{2}$$

where $T\left(\cdot\right)$ is the image processing from RGB to gray.

The encoder E can project the input image to a coarse latent code w, while it is still incapable of faithfully restoring the fine details. To further reconstruct the facial details, a pivot tuning strategy for the generator is followed by the encoder to enhance the identity-consistent reconstruction.

Pivot Tuning Strategy. Pivot tuning Inversion [47] is a method used to improve the inversion quality of GAN-based models. We fix the pre-trained encoder E and further fine-tune the generator G_θ by minimizing the above L_{gray} that measures the difference between the gray-scale of synthesized images and the input gray images. Besides, an ID loss is applied to encourage the extracted features of input images and synthesized images to be as similar as possible. Specifically, the loss function is defined as follows,

$$L_{\text{ID}} = \frac{1}{N} \sum_{i}^{N} \left[1 - \left\langle F\left(G_{\hat{\theta}}\left(w, \boldsymbol{\xi}_i\right)\right), F\left(\mathcal{I}_{\text{gray}}\right)\right\rangle\right], \tag{3}$$

where $F\left(\cdot\right)$ is a pre-trained ArcFace [14] model to extract identity features, and i indicates the i-th view direction of the total N views. Therefore, the fine-tuning procedure of the generator can be expressed as,

$$\theta^{*} = \underset{\theta}{\text{argmin}} \mathcal{L}(G_\theta, \mathcal{I}_{\text{gray}}), \tag{4}$$

where $\mathcal{L}(G_\theta, \mathcal{I}_{\text{gray}}) = L_{\text{gray}} + L_{\text{ID}}$. The colorized image $\mathcal{I}_{\text{color}}$ could be generated from the fine-tued generator:

$$\mathcal{I}_{\text{color}} = G_{\theta^{*}}(w, \boldsymbol{\xi}). \tag{5}$$

3.3. Style Transfer for Historical Image

Inspired by 2D methods [19], we leverage the CLIP model to further fine-tune the 3D-aware generators for high-quality text-guided style transfer. Given the predicted code w and an optimized generator $G_{\theta^{*}}$ after pivot tuning, we continue to train the $G_{\theta^{*}}$ to synthesize the target style images with a target prompt under the supervision of CLIP loss. CLIP loss contains two parts, global loss and directional loss. Global loss measures the similarity between the target text and the style images, which can be presented as,

$$L_{\text{global}} = 1 - \frac{T_{\text{tgt}} \cdot \mathcal{I}_{\text{style}}}{\|T_{\text{tgt}}\| \|\mathcal{I}_{\text{style}}\|}, \tag{6}$$

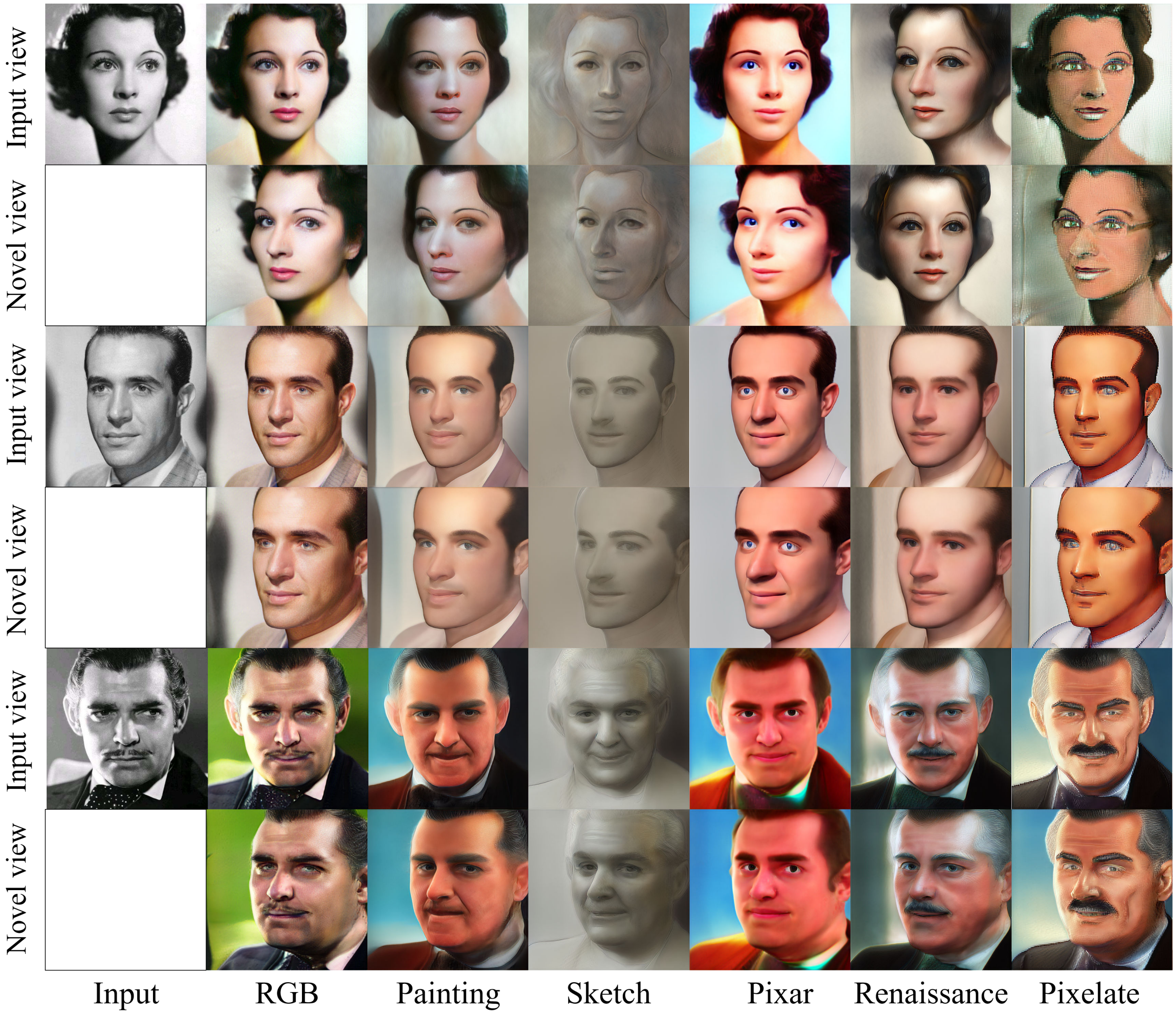

Figure 2. **Qualitative results**. Our HiStyle can restore the color of historic portraits and further transfer the image into diverse styles with multi-views.

where T_{tgt} and I_{style} indicate the feature of the target text and synthesized style image embedded by a CLIP encoder in this section. Another directional loss is to align the changing directions of texts and images which are presented as,

$$L_{\text{direction}} = 1 - \frac{\Delta T \cdot \Delta \mathcal{I}}{\|\Delta T\| \|\Delta \mathcal{I}\|}, \quad (7)$$

where ΔT and $\Delta \mathcal{I}$ are

$$\Delta T = T_{\text{tgt}} - T_{\text{src}}, \quad \Delta \mathcal{I} = \mathcal{I}_{\text{style}} - \mathcal{I}_{\text{color}}. \quad (8)$$

We minimize the CLIP loss as,

$$L_{\text{CLIP}} = \lambda_{\text{global}} L_{\text{global}} + \lambda_{direction} L_{\text{direction}}. \quad (9)$$

However, the text guidance leads to a fixed and specific target of stylized images which can easily cause the mode collapse of the fine-tuned generator and undermine the diversity of style images. To realize the problem of diversity, we propose to take specific style images for each identity as training data. The style data add noise to the direction of fine-tuning procedure for the generator, resulting in diverse characters among different identities. We adopt perpetual loss [26] and directional clip loss as the style loss,

$$\mathcal{L}_{\text{perpetual}}(I_{\text{input}}, I_{\text{style}}) = \sum\nolimits_{l=1}^{L} w_l |G_l(I_{\text{input}}) - G_l(I_{\text{style}})|_2, \quad (10)$$

where I_{input} represents the input image, I_{style} represents the style image, G_l represents the feature maps at layer l of a pre-trained VGG encoder.

However, the style image for each style and each identity is hard to collect. Fortunately, there is not a strict requirement for the training style images of each identity, while they are used to only provide the appropriate disturbance for the text guidance. Therefore, the latent diffusion model is leveraged to generate several style images for each input colorized image under the condition of a text prompt. Although the generation of training style images by diffusion model is not satisfactory in terms of stability and quality, they actually empower us to enhance the diversity for the style transfer.

3.4. Data Augmentation

To increase the diversity of the input dataset and facilitate the training of the 3D-aware GAN for style transfer, we employ a data augmentation method based on Latent Diffusion Model [48]. The latent Diffusion Model is based on the Denoising Diffusion Probabilistic Models (DDPM), which employs a forward process $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ to add noise to the latent representation of the source image $\mathbf{x}_0$, and a reverse process $p(\mathbf{x}_{0:T})$ with a parameterized denoising network to gradually denoise to the target latent representation. Both processes are controlled by a set of noise schedule parameters.

In order to obtain a target latent representation $p(\mathbf{x}_0)$ that is approximately equal to the source latent representation $q(\mathbf{x}_0)$, $p(\mathbf{x}_t) \approx q(\mathbf{x}_t)$ is required. Hence, to ensure that the forward process is approximately equal to the reverse process, DDPM is trained by minimizing a weighted evidence lower bound (ELBO):

$$\mathcal{L}_{\text{Diff}}(\phi, \mathbf{x}) = \mathbb{E}\left[w(t)\|\epsilon_\phi(\alpha_t \mathbf{x} + \sigma_t \epsilon; t) - \epsilon\|_2^2\right], \quad (11)$$

where w(t) is a weighting function, time step $t \sim \mathcal{U}(0,1)$, random noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and ϕ denotes the parameters of the denoising network.

Based on the latent diffusion model, stable random variables are used to generate multiple stylized images from a single input photo, which can serve as additional training samples for the GAN. Given an input image I, we first apply a random affine transformation to produce a slightly modified version of I. We then use stable diffusion to generate K additional images $J_1, J_2, ..., J_K$, each with a different degree of stylization.

$$J_k = I + \sigma_k * X_k, \quad (12)$$

where X_k is a stable random variable with location parameter 0 and scale parameter 1, and σ_k is a scalar controlling the degree of stylization. We set the scale parameter to 1.5 and vary the shape parameter to generate different degrees of stylization. Specifically, we use a shape parameter of 1 for mild stylization, 0.5 for moderate stylization, and 0.3 for strong stylization.

Method	Baseline	HiStyle
photo realism ↑	3.4	**3.6**
text correspondence ↑	**4.1**	3.4
3D consistency ↑	2.5	**4.8**

Table 1. **User Study** conducted with 50 participants, rating the samples on a scale of 1 to 5, with higher scores indicating better quality. The scale ranges from 1 (worst) to 5 (best).

Diffusion model enpowers us to generate various 2D style images to fine-tune the 3D generator, while it also suffers the problem of unstable generation on the aspect of style and pose. On the one hand, the extreme bias on the style of augment data may severely twist the direction when fine-tuning the 3D-aware generator guided by the text prompt. As a consequence, the stylized images generated by the 3D generator shows inconsistent to the text prompts and low image quality. On the other hand, the pose gap between the augment data and the input colorized image can lead to the shape deformation when adapting the domain of a 3D-aware generator. To address the problems of style and pose, the ControlNet is introduced to constrain the augment image for the diffusion model.

By augmenting the input photo I into multiple stylized images $J_1, J_2, ..., J_K$, the dataset is effectively expanded and increases the diversity of the training samples. We then use these augmented images to guide the 3D-aware GAN to adapt to different styles during training. This enables the GAN to learn a more comprehensive representation of the style space, and produce more diverse and realistic stylized portraits.

4. Experiments

4.1. Qualitative results

Colorized Results. In this section, we evaluate the quality of colorized images. Our colorization model is applied to colorize legacy black-and-white photographs. Fig. 3 shows sample results along with manual colorization results by human experts. Compared with the artificial colorization of artists, our HiStyle achieves comparable realism and naturalness of face color.

Stylized Results. In this section, we discuss the experiments conducted to evaluate the quality of stylized images. We display a wide range of text-driven stylized results of our HiStyle in Fig. 2. our model can imagine the realistic color of historical gray images, and moreover enable us to synthesize multi-view consistent images in various style domains.

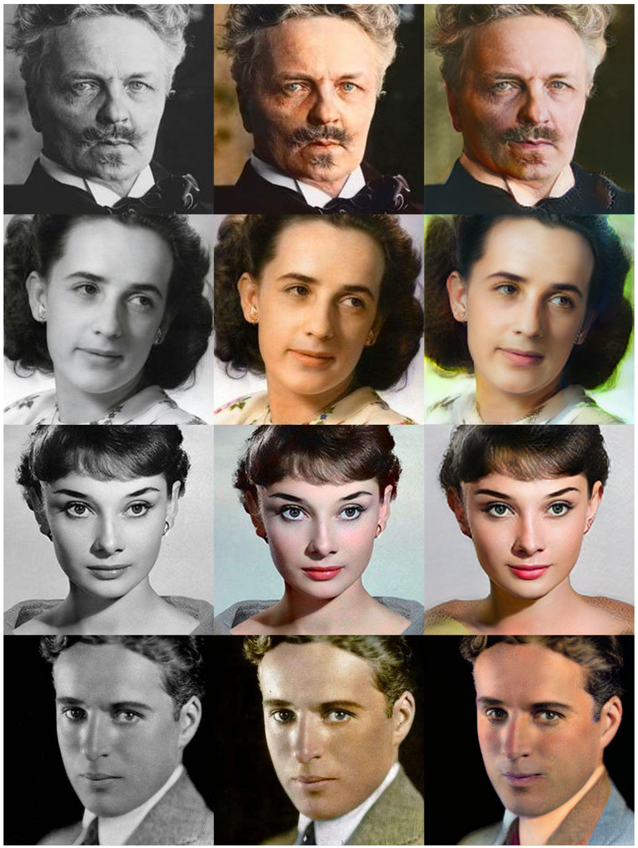

Figure 3. **Comparison with manual colorization.** Our colorization model can achieve comparable results with those colorized by professional artists.

4.2. Quantitative results

User study. We conducted a user study involving 50 historical gray images that were restored into RGB images and further stylized using text-to-image models. The aim of the study was to evaluate the perceptual quality of the generated samples. Participants were asked to rate the samples on a scale of 1 to 5 based on the following criteria:

Photo Realism: Does the sample align with your impression of a realistic person? *Text Correspondence:* Does the generated sample accurately reflect the semantics of the target text? *3D Consistency:* Are the samples consistent from various viewpoints? The ratings provided by the participants were used to quantify their opinions.

As shown in Table 1, our results demonstrate the superior quality, high diversity, and strong text-image correspondence achieved by our proposed method compared to the baselines (2D-based method). This user study provides valuable insights into the perceptual evaluation of the generated samples and highlights the strengths of our approach.

3D consistency In this section, several experiments are conducted for the evaluation of 3D consistency.

Method	ID similarity ↑
EG3D [8]	0.80
HiStyle (Colorization)	0.79
HiStyle (Stylization)	0.68

Table 2. **Quantitative results of 3D consistency**. Our model demonstrates comparable performance after colorization. Despite incorporating additional manipulations, our model maintains its effectiveness in terms of depth consistency and identity consistency across multiple views. However, since the ArcFace [14] pretrained model is not adaptable to style domains, the results for identity similarity in stylized images are just for reference.

The evaluation of 3D consistency in 3D-aware GANs often relies on the ArcFace [14] cosine similarity, which measures facial identity similarity between multi-views. To evaluate facial identity consistency, a similar pre-processing step is followed as in the evaluation of depth consistency. To assess the facial identity consistency of EG3D as a baseline, high-resolution colorized images of size 512×512 are generated from two randomly selected side views, and the cosine similarity between these views is measured. For HiStyle, we apply our model to colorize and stylize 100 collected historic portraits, generating a novel view. The cosine similarity between the input view and the novel view is similarly measured as EG3D. However, it should be noted that the ArcFace model, being pre-trained on the FFHQ dataset, is not designed to evaluate the identity similarity of style images. Therefore, the measurement of stylized portraits is not accurate, just provided for reference.

As demonstrated in Table 2 and illustrated in Figure 4, the colorization process maintains good 3D consistency, as observed in the case of EG3D. Multiple examples with various views and geometric shapes are presented to provide a visual representation of the 3D consistency achieved

4.3. Ablation Study

In this study, we evaluate the effectiveness of each individual component, *i.e*. Data Augmentation, RGB-to-Gray loss, encoder and pivot tuning in enhancing the color of images.

Data Augmentation. To show the effect of our data augmentation, an experiment is conducted that compares the results with the style data augmentation of the latent diffusion model and the results without it. As shown in Fig. 5, the stylization results with LDM augmentation show more significant style changes and more consistency with the input style prompt. Besides, the diversity among different identities also shows the superiority of the model with LDM. For example, the eyes color of all the results without LDM are blue, while the results with LDM show diverse eye colors

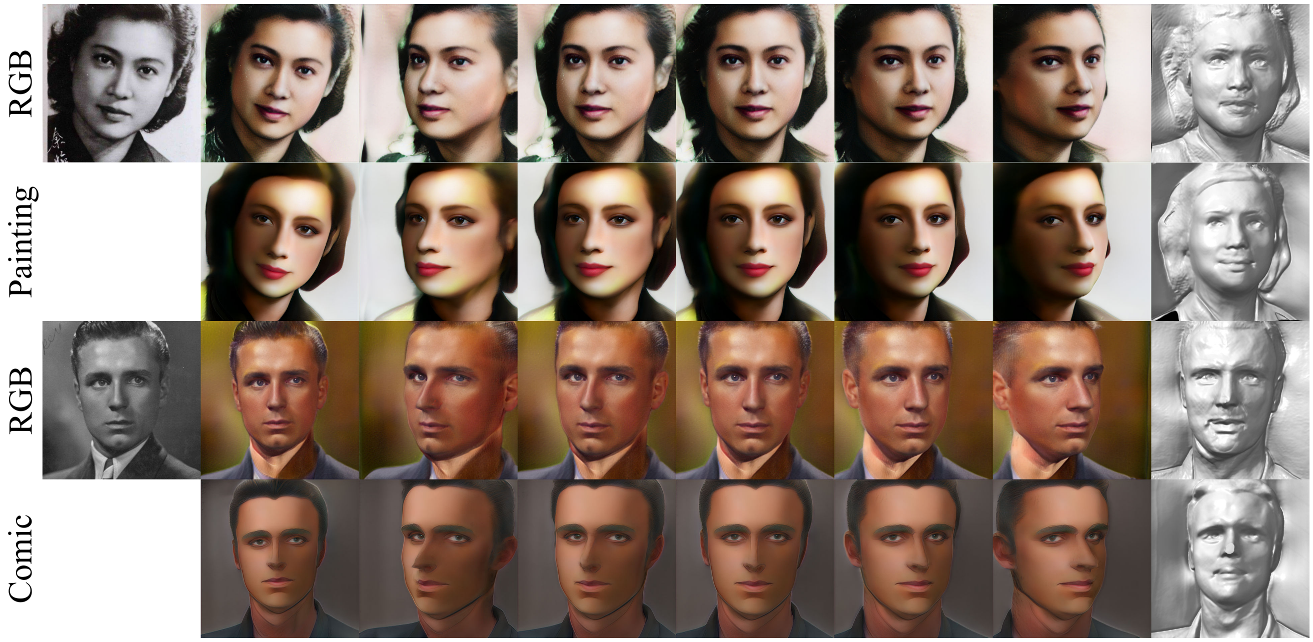

Figure 4. **Results of 3D consistency**. When presented with an input historical portrait, our model preserves excellent 3D consistency in the colorized and stylized outputs. The generated images exhibit consistent geometry while showcasing minimal variations in appearance across a collection of five-view images. Additionally, the inclusion of mesh results further demonstrates the high-quality geometry produced by our approach. This highlights the ability of our model to maintain accurate 3D representations throughout the colorization and stylization process.

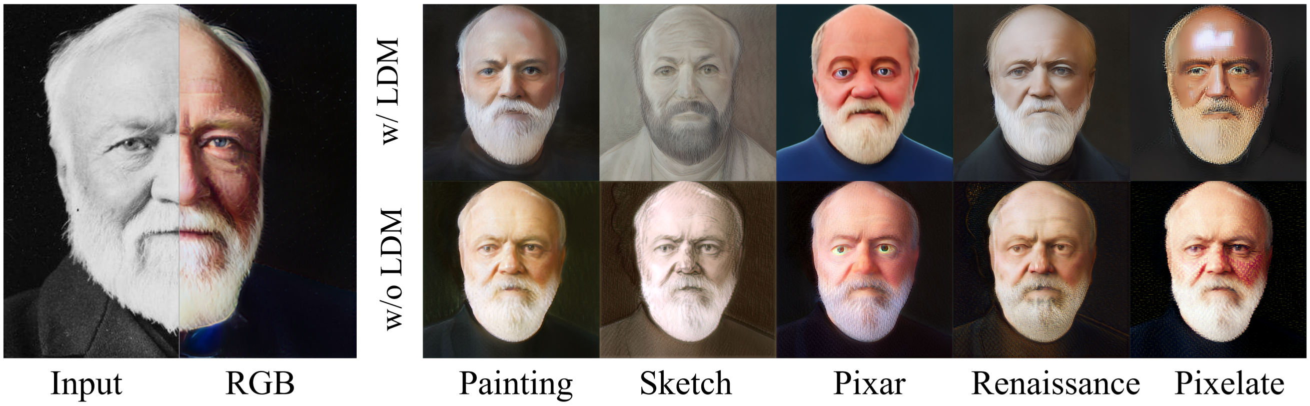

Figure 5. **Ablation study** on data augmentation. When data augmentation by LDM is not applied, the stylized results demonstrate minimal diversity and exhibit similar appearances, lacking distinct variations in style. However, with the integration of data augmentation, the results exhibit a multitude of vibrant and visually appealing styles, showcasing a wide range of variations and enhancing the overall richness of the generated images. The inclusion of data augmentation by LDM significantly contributes to the generation of diverse and captivating stylized outputs.

in different samples.

RGB-to-Gray Loss. As shown in Fig. 6, the results without RGB-to-Gray loss show a colorless face close to the gray image. Our results demonstrate that the RGB-to-Gray loss is significantly effective in encouraging the generator to produce images with more realistic colors.

Initial Encoder. Fig. 6 indicates that incorporating an initial encoder also significantly improves the colorfulness of the generated images. Specifically, the encoder provides an initial point for the optimization process, avoiding the unstable search in latent space, meanwhile reducing the number of optimization iterations.

Pivot Tuning. Despite the nature of color achieved by the encoder and RGB-to-Gray loss, the results without fine-

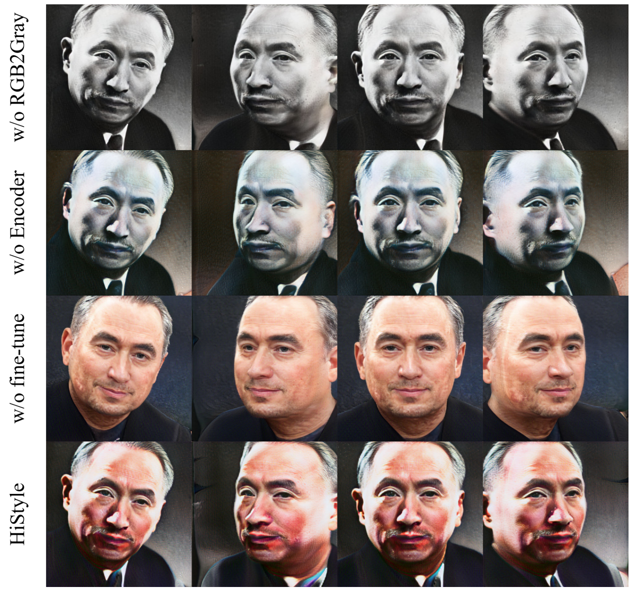

Figure 6. **Ablation study** on enhancing the color of images. Without RGB-to-Gray loss or initial encoder, the results show a colorless appearance close to the gray images, while without fine-tuning, the results show inconsistent identity and lack of details.

tuning show inconsistent identities and a lack of details in Fig. 6. It highlights the significance of fine-tuning in photorealism.

5. Limitations and Future Work

Despite the promising results achieved by our method, there are still limitations that should be addressed in future work. Firstly, our method is limited to restoring the color of the head in historical portraits, and cannot be applied to restore the color of the whole body. This is due to the fact that the dataset used only includes images of historical portraits that depict the head and shoulders of the subjects. Besides, the full pipeline is a lengthy optimization procedure which prevents the method from real-time applications. Further research is needed to address these limitations and extend the applicability of our method to a wider range of historical images.

6. Conclusion

In this project, we have presented a novel approach for reinventing historical portraits through colorization, stylization, and 3D lifting. Leveraging the power of pre-trained text-image models and recent advancements in 3D-aware image representation, our technique transfers 2D historical images into vibrant 3D representations. This is achieved by employing GAN inversion to map gray historical portraits into the latent space of a 3D generator, enabling the restoration of lost color information and reconstruction of the original portrait's geometric shape.

To further enhance the stylization aspect, we fine-tuned the 3D generator using guidance from the CLIP model, allowing for style transfer from a source domain to a target domain. Overcoming challenges related to data diversity and generated image variety in the style domain, we introduced the concept of augmenting target style data using latent diffusion models after colorization. This innovative approach, called HiStyle, successfully resolves issues related to diversity loss and encompasses all three dimensions of the historic reinvention task.

Our contributions encompass the proposal of a novel network architecture, the integration of diffusion models into 3D-aware GANs for style data augmentation, and the introduction of the gray-inversion loss for GAN inversion of gray images. Through extensive experimentation, we have demonstrated the effectiveness of our approach in transferring historical portraits into 3D colorful images with diverse styles. This significant advancement surpasses existing methods and establishes new possibilities in the realm of reinventing historical portraits.

7. Acknowledgement

This project derives from the existing techniques and utilizes them in a field that is relevant to the everyday life of individuals. The choice of this topic is due to my personal interest in history and my belief that history is an important topic to be taught. Such a topic would also fit the functions of this project quite well, since historical photos are often of quite different styles of modern photographs and would be fitting for style transfers. Finally, the choice of this topic is something that could be used for the betterment of historical education, making distant figures feel closer to regular people. During the work on this project, there were many concepts and processes that needed to be learned and utilized immediately without institutionalized education from the very beginning. I would like to thank Mr. Chen Zhuo and Prof. Li Zhu (the director of the NSF Center for Big Learning at UMKC, and the Associate Editor-in-Chief for IEEE Transactions on Circuits and Systems for Video Technology) for helping me understand the processes involved in the creation of such a systems. Such a project would not be possible without their diligent tutelage, and the knowledge gained from their guidance has been invaluable.

References

[1] Victoria Fernández Abrevaya, Stefanie Wuhrer, and Edmond Boyer. Multilinear autoencoder for 3d face model learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2018. 3

[2] Aibek Alanov, Vadim Titov, and Dmitry P Vetrov. Hyperdomainnet: Universal domain adaptation for generative adver-

sarial networks. *Advances in Neural Information Processing Systems*, 35:29414–29426, 2022. 2, 3

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 3

[4] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018. 3

[5] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016. 3

[6] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 3

[7] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. Unsupervised diverse colorization via generative adversarial networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pages 151–166. Springer, 2017. 2

[8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16123–16133. 3, 7

[9] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809, 2021. 3

[10] Zhuo Chen, Xudong Xu, Yichao Yan, Ye Pan, Wenhan Zhu, Wayne Wu, Bo Dai, and Xiaokang Yang. Hyperstyle3d: Text-guided 3d portrait stylization via hypernetworks. *arXiv preprint arXiv:2304.09463*, 2023. 2, 3

[11] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *Proceedings of the IEEE international conference on computer vision*, pages 415–423, 2015. 1, 2

[12] Yuanzheng Ci, Xinzhu Ma, Zhihui Wang, Haojie Li, and Zhongxuan Luo. User-guided deep anime line art colorization with conditional adversarial networks. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1536–1544, 2018. 2

[13] Hang Dai, Nick Pears, William AP Smith, and Christian Duncan. A 3d morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE international conference on computer vision*, pages 3085–3093, 2017. 3

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699. 4, 7

[15] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, pages 10673–10683, 2022. 3

[16] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 4

[17] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE international conference on computer vision*, pages 567–575, 2015. 1, 2

[18] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *CVPR*, pages 14304–14313, 2021. 3

[19] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, pages 1–13, 2022. 2, 3, 4

[20] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 1, 2

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

[22] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *ICLR*, 2021. 3

[23] Fangzhou Han, Shuquan Ye, Mingming He, Menglei Chai, and Jing Liao. Exemplar-based 3d portrait stylization. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 1

[24] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 1, 2

[25] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016. 2

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 1, 2, 5

[27] Yucheol Jung, Wonjong Jang, Soongjin Kim, Jiaolong Yang, Xin Tong, and Seungyong Lee. Deep deformable 3d caricatures with learned shape control. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 1

[28] Leila Kiani, Masoud Saeed, and Hossein Nezamabadi-pour. Image colorization using generative adversarial networks and transfer learning. In *2020 International Conference on Machine Vision and Image Processing (MVIP)*, pages 1–6. IEEE, 2020. 2

[29] Gwanghyun Kim and Se Young Chun. Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model. *arXiv preprint arXiv:2211.16374*, 2022. 3

[30] Adam R Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Soňa Mokrá, and Danilo J Rezende. Nerf-vae: A geometry aware 3d scene generative model. *arXiv preprint arXiv:2104.00587*, 2021. 3

[31] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 577–593. Springer, 2016. 1, 2

[32] Kyle Lennon, Katharina Fransen, Alexander O'Brien, Yumeng Cao, Matthew Beveridge, Yamin Arefeen, Nikhil Singh, and Iddo Drori. Image2lego: Customized lego set generation from images. *arXiv preprint arXiv:2108.08477*, 2021. 1

[33] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NIPS*, 2017. 1, 2

[34] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *ECCV*, 2018. 1, 2

[35] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 99–106. 3

[36] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[37] Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi. Image colorization using generative adversarial networks. In *Articulated Motion and Deformable Objects: 10th International Conference, AMDO 2018, Palma de Mallorca, Spain, July 12-13, 2018, Proceedings 10*, pages 85–94. Springer, 2018. 2

[38] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. *arXiv preprint arXiv:2103.17269*, 2021. 3

[39] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, pages 11453–11464, 2021. 3

[40] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, pages 13503–13513, 2022. 3

[41] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *NIPS*, 2021. 3

[42] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *CVPR*, pages 2085–2094, 2021. 2

[43] Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4142–4160, 2020. 3

[44] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10934–10943, 2019. 3

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[46] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. In *ICML*, 2021. 3

[47] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 4

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 6

[49] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NIPS*, 2020. 3

[50] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022. 3

[51] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7968–7977, 2020. 2

[52] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019. 3

[53] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3d morphable face models from images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3361–3371, 2021. 3

[54] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 4

[55] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019. 3

[56] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):157–171, 2019. 3

[57] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, pages 6924–6932, 2017. 1, 2

[58] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. *arXiv preprint arXiv:2212.06135*, 2022. 3

[59] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255*, 2022. 3

[60] Yi Xiao, Peiyao Zhou, Yan Zheng, and Chi-Sing Leung. Interactive deep colorization using simultaneous global and local inputs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1887–1891. IEEE, 2019. 2

[61] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. In *NIPS*, 2021. 3

[62] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *CVPR*, 2022. 3

[63] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 1, 2

[64] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017. 2

[65] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 3