3DToonify: Creating Your High-Fidelity 3D Stylized Avatar Easily from 2D Portrait Images

Yifang Men[1]*, Hanxi Liu[2]*, Yuan Yao[1], Miaomiao Cui[1], Xuansong Xie[1], Zhouhui Lian[2]†
[1]Institute for Intelligent Computing, Alibaba Group
[2]Wangxuan Institute of Computer Technology, Peking University, China

(a) Input video (b) 3D style adaption (c) View-consistent renderings in different styles

Figure 1. Given a set of RGB portrait images captured by a monocular camera, our method can learn a photorealistic representation in neural implicit fields, and transfer it to artistic ones with underlying 3D structures changed. Multiple stylized results can be rendered from arbitrary novel viewpoints with consistent geometry and texture.

Abstract

Visual content creation has aroused a surge of interest given its applications in mobile photography and AR/VR. Portrait style transfer and 3D recovery from monocular images as two representative tasks have so far evolved independently. In this paper, we make a connection between the two, and tackle the challenging task of 3D portrait stylization – modeling high-fidelity 3D stylized avatars from captured 2D portrait images. However, naively combining the techniques from the two isolated areas may suffer from either inadequate stylization or absence of 3D assets. To this end, we propose 3DToonify, a new framework that introduces a progressive training scheme to achieve 3D style adaption on spatial neural representation (SNR). SNR is constructed with implicit fields and they are dynamically optimized by the progressive training scheme, which consists of three stages: guided prior learning, deformable geometry adaption and explicit texture adaption. In this way, stylized geometry and texture are learned in SNR in an explicit and structured way with only a single stylized exemplar needed. Moreover, our method obtains style-adaptive underlying structures (i.e., deformable geometry and exaggerated texture) and view-consistent stylized avatar rendering from arbitrary novel viewpoints. Both qualitative and quantitative experiments have been conducted to demonstrate the effectiveness and superiority of our method for automatically generating exemplar-guided 3D stylized avatars.

1. Introduction

Portrait style transfer [30, 53] aims to transform real face images into artistic 2D portraits in desired visual styles while maintaining personal identity. However, given a sequence of portrait images captured from different viewpoints, existing portrait style transfer methods are typically

*Denotes equal contribution.
†Corresponding author. E-mail: lianzhouhui@pku.edu.cn.
This work was partially supported by National Natural Science Foundation of China (Grant No.: 62372015).

only effective for limited forward-facing photos and fails to maintain view consistency in 3D space. Essentially, existing methods only learn a style transfer between 2D features, and have no sense to 3D representations built on real-world objects. What if we can construct and stylize underlying 3D structures from captured 2D portrait images? See Figure 1 for an example. When stylized with 3D structures (i.e., geometry and texture), we can easily render view-free stylized portraits with 3D consistency and robust artistic results. This capacity will extremely facilitate the 3D content creation process which often requires large amounts of time and special expertise, and make it accessible to a variety of novice users. As shown in Figure 1, this paper aims to address the challenging task of generating high-fidelity 3D avatar from a portrait video by following the style of a given exemplar image. We refer this task as *3D portrait stylization* – a marriage between portrait style transfer and 3D recovery from monocular images.

The naïve solution to the task mentioned above is directly combining existing methods of 2D portrait stylization with 3D reconstruction, i.e., learning 3D representations such as voxels [36], primitives [26] or occupancy fields [31] directly from stylized portrait images. However, it is less effective due to the biased image manifold built by 2D portrait stylization, making the representation learning be ill-posed with highly-biased visible views. Recently, neural radiance field (NeRF) [4, 18, 32, 33, 40, 49] has made great progress due to its advanced ability to achieve photo-realistic novel view synthesis with sparse input views. Some previous attempts [8, 34, 35, 48, 58] also combine NeRF with image-based [11] or text-driven [42] neural style transfer to generate novel views of stylized 3D scenes or avatars. Recently, a series of new works have started to focus on 3D stylized avatar generation. Some methods [7, 15, 23, 25, 27, 41, 50, 56] exploit the great potential of 2D text-to-image diffusion models [44–46] to generate 3D cartoonish avatars according to a given text prompt. Others [2, 51, 55, 57] build on 3D generative models [6, 38] to bridge the gap between the real space and the target domain, and generate avatars with certain styles under a sampled latent vector. However, all these methods either can not achieve high-fidelity personalized 3D portrait stylization with user-specific identities and styles, or fail to generate fine-grained full-head avatars that support view-consistent rendering from arbitrary viewpoints.

To address the aforementioned challenges, we draw inspiration from domain adaption on 2D features [10, 30], and introduce a progressive training scheme to achieve 3D style adaption on spatial neural representation (SNR). The key insights of this design are twofold. First, it is hard to directly learn an accurate 3D representation field from stylized portraits with few-shot inconsistent 2D views, but easier to learn a photorealistic field as a prior and adapt it to target style fields with transfer learning. Second, learning spatial representation with disentangled surface and texture allows for flexible geometry deformation and texture adaption, leading to more diverse and fine-grained style editing. To this end, we construct SNR with neural implicit fields and dynamically optimize its subfields with a progressive training scheme. This scheme includes the following three stages: *prior learning* to obtain an accurate human reconstruction, *geometry adaption* to produce inherently exaggerated deformation, and *texture adaption* to realize artistic albedo decomposition. Eventually, the 2D portraits are converted to stylized SNR, and explicit 3D assets can be easily extracted with disentangled 3D structures. In summary, our contributions are threefold:

- We present a new method that adopts neural implicit fields to address the challenging task of generating high-fidelity 3D avatar from a portrait video by following the style of a given exemplar image. Stylized results can be rendered under arbitrary novel viewpoints with consistent geometry and texture.
- We introduce an elegant network of spatial neural representation to model common attributes over the 3D space. This design allows for disentangled geometry and texture adaption, achieving more flexible and fine-grained 3D stylization results.
- We propose a novel progressive training scheme of 3D style adaption. Cooperated with the delicately-designed spatial neural network, it enables learning realistic 3D cartoon avatars with deformed geometry and stylized texture.

2. Related Work

2D Portrait Stylization. In the deep neural network based portrait stylization, there are two types of approaches, i.e., image-to-image translation and StyleGAN based translation. Methods [21, 24] conduct face-to-cartoon translation by adopting the framework of cycleGAN [60]. Nevertheless, training such methods requires extensive data and may still generate unstable results. StyleGAN [16, 17] has become a popular alternative for portrait stylization due to its strong capacity for latent inversion and style control. [30] proposes a calibration framework to adapt the original training distribution for fine-grained translation. [53] leverages the mid- and high-resolution layers of StyleGAN to render high-quality artistic portraits based on the multi-scale content features to better preserve details. Although high-quality results have been shown, these methods cannot handle extreme face angle while maintaining cross-view consistency.

Neural Implicit Fields. Recently, neural implicit functions have emerged as an effective representation to model conventional 3D scenes due to its continuous nature. This representation has been successfully adopted to shape model-

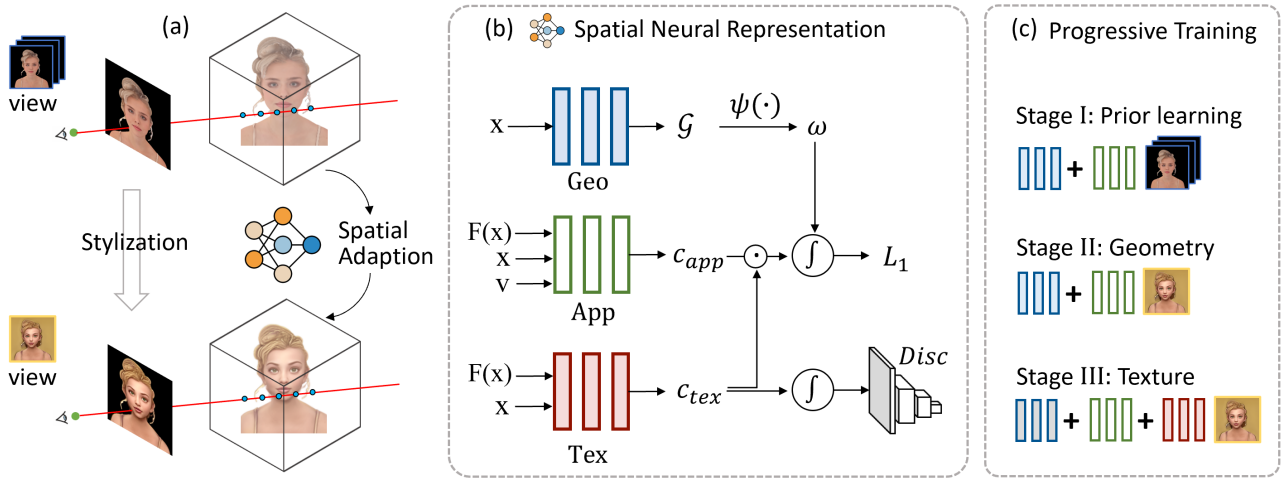


Figure 2. An overview of the proposed framework. Our method first learns a photorealistic field built-upon spatial neural representation (SNR) using dense input views, then transfers this prior representation to artistic ones with few-shot stylized views by adapting underlying 3D structures. SNR is constructed by a geometry field for SDF surface, an appearance field for observed color, and a texture field for albedo color, respectively. The progressive training scheme is adopted to enable SNR to learn about stylized geometry and texture in an explicit and structured manner.

ing [12, 39], novel view synthesis [29, 32] and multi-view 3D reconstruction [49, 54]. The method of Neural Radiance Fields (NeRF) [32], in particular, has attracted significant attention for its ability to achieve photo-realistic novel view synthesis results by utilizing neural implicit functions together with volume rendering. A number of variants have been developed thereafter to fit with different scenarios and requirements, including quality improvement [4], fast rendering [33], dynamic scene capture [40] and generative models [5]. However, NeRF's estimated volume density does not admit accurate surface reconstruction, the recovered 3D geometry is far from satisfactory and can hardly be extracted as explicit materials. Recent works tackle the issue by combining implicit surface functions. [37] represents the surface by occupancy values and shrink the sample region of volume rendering during the optimization. [49] introduces signed distance functions (SDF) to represent the scene and can directly extract the surface as the zero-level set of the SDF with better accuracy.

3D Avatar Stylization. 3D avatar stylization aims to generate stylized 3D avatars whose rendered images captured from different viewpoints match the specific style. Early methods are either mesh-driven [13] or rely on explicit parameterization [47]. More recently, [35, 48] exploit the flexibility of neural radiance field and propose a text-guided stylization approach that manipulates the reconstructed scenes with input text prompts. However, due to the limited expressiveness of natural languages, they can not generate highly-detailed results with arbitrary user-specific styles. Another stream of methods [2, 19, 20, 57] using 3D generative models [6, 38] have extended avatar stylization to 3D-aware domain adaption. However, inherited from their predecessors, these methods can not synthesize full-head avatars in 360°, and perform badly with real-world out-of-domain data. In contrast, our method utilizes the implicit representation to model high-fidelity 3D avatars from captured portrait videos, which allows for superior view consistency and stable stylization.

3. Method Description

Given the short portrait video of a person captured with a monocular camera, we aim to generate the high-fidelity 3D stylized avatar of the person. The person stands still when recording the video. We denote the split frames of the video as $\{I_i | i = 1, \ldots, N\}$, where i is the frame index, N is the number of frames. For each frame, we use COLMAP to obtain the calibrated camera and the method proposed in [28] to extract the foreground human mask.

The overview of the proposed framework is illustrated in Figure 2. 3DToonify aims to learn the stylized human neural field by adapting 3D structures in a progressive training scheme. This scheme is built upon a spatial neural representation, which utilizes disentangled implicit fields to capture the underlying 3D structures such as geometry and texture (Section 3.1). We first leverage the geometric guidance from a multi-view stereo to learn a robust photorealistic representation, acting as a source prior (Section 3.2). Then this prior representation is adapted to the style domain with adaptive geometry deformation (Section 3.3.1) and decomposed albedo colors (Section 3.3.2). In this way, the stylized human avatar field can be constructed by SNR with transformed underlying structures, thus allowing for fully stylized results and 3D consistent rendering in arbitrary viewpoints.

3.1. Spatial neural representation

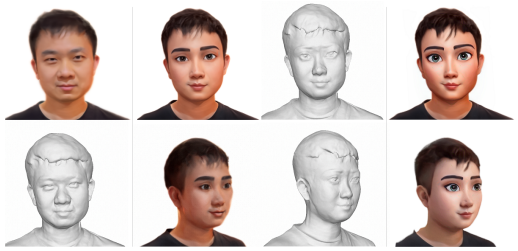
The proposed spatial neural representation (SNR) is based on neural radiance field (NeRF) [32], which can be seen as a continuous 5D function that maps a 3D position $\mathbf{x}$ and a viewing direction $\mathbf{v}$ to an emitted color $\mathbf{c} = (r, g, b)$ and a volume density σ. NeRF is approximated by a multi-layer perceptron (MLP) $F_\theta : (\mathbf{x}, \mathbf{v}) \rightarrow \mathbf{c}, \sigma$. SNR consists of three MLPs F_{geo}, F_{app} and F_{tex}, representing the decomposed fields of geometry, the observed appearance color and the albedo texture color, respectively.

Geometry field learns a function $F_{geo} : \mathbb{R}^3 \rightarrow \mathbb{R}$ that maps a spatial point $\mathbf{x} \in \mathbb{R}^3$ to its signed distance value $\mathcal{G}$ to the object surface. It constructs the underlying object surface by encoding a signed distance function (SDF) of only location $\mathbf{x}$. In order to be compatible with the rendering procedure of the radiance field, a probability function $\psi(.)$ proposed by [49] is used to calculate the point weight w from the signed distance value $\mathcal{G}$, where $\psi(.)$ denotes an unbiased and occlusion-aware approximation. With this implicit SDF representation, the explicit object surface S can be easily extracted by the zero level-set of the $SDF : S = \{\mathbf{x} \in \mathbb{R}^3 | \mathcal{G}(\mathbf{x}) = 0\}$.

Appearance field learns a function $F_{app} : \mathbb{R}^3 \times \mathbb{S}^2 \rightarrow \mathbb{R}^3$ to encode the observed colors $\mathbf{c}_{app}$ associated with the point $\mathbf{x} \in \mathbb{R}^3$ and the view direction $\mathbf{v} \in \mathbb{S}^2$. The feature vectors $F(\mathbf{x})$ derived from F_{geo} are also concatenated as the inputs. To better approximate the appearance colors of the object captured in read-world scenes, F_{app} is introduced as a function of both location and viewing direction, thus allowing learning view-dependent RGB colors for multi-view images. Notably, the learned representation in F_{app} could be degraded into reflection components $\mathbf{s}$, which are caused by illumination and vary with view directions. It will be adaptively changed in the later training stage (see the detailed discussion in Section 3.3.2).

Texture field learns a function $F_{tex} : \mathbb{R}^3 \rightarrow \mathbb{R}$ to encode the albedo color for the texture atlas $\mathbf{c}_{tex}$ associated with only the spatial location $\mathbf{x}$. Similar to F_{app}, feature vectors derived from F_{geo} are concatenated as inputs. We encourage the texture representation to be multi-view consistent by restricting F_{tex} being a function of only $\mathbf{x}$, while allowing the final color $\mathbf{c} = \mathbf{s} \circ \mathbf{c}_{tex}$ to be view-dependent to satisfy different view observations, where $\circ$ denotes element-wise multiplication. With the nature of view-independent representation of F_{tex}, explicit textures can be obtained by accumulating the volume albedo colors.

The proposed geometry field and texture field are formulated in a view-independent function, once being effectively learned, they can express spatial attributes shared by the entire 3D space. This enables editable 3D structures with only few-shot stylized views needed in the later adaption process.



(a) Stage I (b) Stage II (c) Stage III

Figure 3. Visualized results in stage I, II, III.

3.2. MVS guided prior learning

In this module, we learn the photorealistic representation as a prior for the later 3D style adaption. Due to the complexity of real-world captures caused by illumination, object materials, etc., the reconstructed results can easily suffer from noisy surfaces and irregular holes. Observing that the geometry directly extracted by multi-view stereo (MVS) methods are generally accurate with only local noises, we propose to integrate the depth information estimated by MVS as a geometric guidance for surface reconstruction.

Accumulated depth guidance. Volume rendering has been proven effective to enable robust supervision using 2D image observations. Following this, we render the depth map with K points along the emitted ray and use the corresponding 2D depth value for supervision. The ray can be parametrized as $\boldsymbol{r}(i) = \boldsymbol{o} + d_i\mathbf{v}$, where $\boldsymbol{o}$ is the center of the camera and $\mathbf{v}$ is the direction of the ray. The depth $\hat{D}(r)$ from the geometry field can be computed by:

$$\hat{D}(r) = \sum_{i=1}^{K} (T_i \alpha_i d_i), \tag{1}$$

where T_i is the accumulated transmittance defined by $\Pi_{j=1}^{i+1}(1 - \alpha_j)$, and α_j denotes the discrete opacity value computed by $\alpha_j = max(\frac{\Phi_s(s_i) - \Phi_s(s_i+1)}{\Phi_s(s_i)}, 0)$, in which Φ is the cumulative distribution of logistic distribution. More details about conversion from the SDF distance to the opacity can be found in NeuS [49]. For a batched training ray $r \in R$, the accumulated depth loss can be formulated as:

$$L_{depth} = \sum_{r \in R} ||M(r)(\hat{D}(r) - D(r))||_1, \tag{2}$$

where $M(r) \in \{0, 1\}$ is the object mask value and $D(r)$ is the supervised depth value.

Depth-sampled surface guidance. Except for the depth constraint on spatial accumulated points, we also leverage points sampled from the depth image I_D to guide the construction of the SDF surface. The surface loss encourages these sampled 3D points being close to the object surface and L_{sur} can be formulated as:

$$L_{sur} = \sum_{\mathbf{x}_d \in I_D} ||F_{geo}(\mathbf{x}_d)||_1. \tag{3}$$

Training. Given a set of portrait images and their camera parameters, we train the architecture with the geometry field and the appearance field using the following loss function:

$$L_{prior} = L_{color} + \lambda_{mvs} L_{mvs} + \lambda_{mask} L_{mask} + \lambda_{reg} L_{reg}, \tag{4}$$

where λ denotes the weight of each corresponding loss. The MVS guided loss is computed as $L_{mvs} = L_{depth} + L_{sur}$. The color reconstruction loss L_{color} is calculated as the distance between the accumulated color $\hat{C}(r)$ and the observed color $C(r)$ of I :

$$L_{color} = \sum_{r \in R} ||M(r)(\hat{C}(r) - C(r))||_1, \tag{5}$$

where $\hat{C}(r)$ can be computed by $\sum_{i=1}^{K}(T_i \alpha_i c_i)$, and c_i denotes the volumetric color produced by the appearance field F_{app}. To focus on human reconstruction, we also define a mask term with the binary cross entropy loss:

$$L_{mask} = BCE(\hat{M}(r), M(r)), \tag{6}$$

where $\hat{M}(r) = \sum_{i=1}^{K}(T_i \alpha_i)$ is the density accumulation along the ray. The Eikonal loss [12] used to regularize the SDF values is defined as

$$L_{reg} = \sum_{k} ||\nabla_{\mathbf{p_k}} F_{geo}(\mathbf{x_k}) - 1||_2^2. \tag{7}$$

Visualized results of this stage are shown in Figure 3 (a). Not only the radiance field with accumulated color is learned, but also the inherent geometry can be accurately decomposed. The high-quality reconstruction learned in this stage also paves the way for the next stage of style adaption with few-shot 2D stylized portraits.

3.3. Spatial representation adaption

With the constructed photorealistic representation, we then transform it to the style domain by progressively adapting the underlying 3D structures. We first adaptively learn the faithful deformed geometry without the interference of the albedo texture module, and then decompose albedo colors from observed ones with fixed geometric structures. This enables effective 3D structure disentanglement with more accurate surface and clearer texture.

3.3.1 Geometry adaption

In this stage, we utilize a number of stylized 2D portrait images I_t derived from exsting 2D portrait stylization methods [30, 53] to fine-tune the geometry field F_{geo} and the appearance field F_{app}. The spatial-shared geometry will be adaptively transformed in F_{geo} and the observed colors varying with views will be modeled in F_{app}, enabling the network focusing on geometry adaption. During training, the pixel color of I_t is used as the observed color to guide the accumulated volume colors:

$$L_{color} = \sum_{r \in R} ||M(r)(C(r) - C_t(r))||_1, \tag{8}$$

where $C(r)$ is computed by the volumetric color from F_{app} and the converted opacity from F_{geo}. The total training loss is formulated as:

$$L_{geo} = L_{color} + \lambda_{mask} L_{mask} + \lambda_{reg} L_{reg}. \tag{9}$$

As shown in Figure 3 (b), the spatial deformed geometry can be extracted from F_{geo}. However, rendering results are 3D-inconsistent with obvious artifacts in side-view renderings, since only few-shot 2D stylizations of the frontal views are provided for style adaption and the view-dependent function F_{app} trivially fits these views.

3.3.2 Albedo texture adaption and optimization

In this stage, we aim to learn the spatial-shared texture field F_{tex} by decomposing the albedo colors from the appearance ones. Specifically, we insert F_{tex} as a view-independent texture field and jointly optimize F_{tex} and F_{app}. In this way, view-consistent colors can be effectively decomposed from the total appearance and the remaining components in F_{app} are regarded as view-dependent reflections. The final color are computed by $\tilde{c}_i = s \circ c'_i$, where c'_i is the albedo color from F_{tex} and s is the degraded reflection from F_{app} for spatial points. Then we can obtain the final accumulated color by

$$\tilde{C}(r) = \sum_{i=1}^{K}(T_i \alpha_i \tilde{c}_i). \tag{10}$$

To further ensure effective albedo color decomposition, a discriminator D is introduced to encourage $\tilde{C}(r)$ satisfying the approximate distribution of palette colors of I_t. With κ as a posterize filter, the patch color $\kappa(C_t(p))$ of I_t is fed into D as a real sample, and the reconstructed color $\tilde{C}(p)$ from F_{tex} is fed into D as a fake sample, where p is the set of rays for image pixels in a patch. We define the discrimination loss L_{ds} to penalize for distance between the distribution of $C(p)$ and $\tilde{C}(p)$ as:

$$\begin{aligned} L_{ds} =& \mathbb{E}_{p \sim \{I_t^i\}}[log(D(\kappa(C_t(p))))] + \\ & \mathbb{E}_{p \sim \{I_t^i\}}[log(1 - D(\tilde{C}(p)))]. \end{aligned} \tag{11}$$

To keep the learned geometry stay faithful to the given style, we fix F_{geo} and train $\{F_{app}, F_{tex}\}$ with the training loss as follows:

$$L_{tex} = L_{color} + \lambda_{mask} L_{mask} + \lambda_{reg} L_{reg} + \lambda_{ds} L_{ds}, \tag{12}$$

where L_{color} denotes the distance between the final accumulated color $\tilde{C}(r)$ and the observed stylized color $C_t(r)$:

$$L_{color} = \sum_{r \in R} ||M(r)(\tilde{C}(r) - C_t(r))||_1. \tag{13}$$

(a) Input (b) Stylized results in novel views (c) Exported meshes

Figure 4. Stylized results in novel views and corresponding exported meshes.

We show rendering results of this stage in Figure 3 (c), demonstrating their 3D consistency in multi-view setting. Thanks to the spatial-shared colors learned in the view-independent F_{tex}, the albedo texture can be seamlessly extracted and further enhanced in an explicit manner.

4. Experimental Results

Implementation details. Our network architecture consists of three modules: the signed distance function F_{geo}, the appearance function F_{app} and the texture function F_{tex}, which are modeled by three MLPs with 8, 6, 6 hidden layers, respectively. Positional encoding [32] and sphere initialization [3] are also applied similar to [49]. For the depth priors, we adopt the OpenMVS method [1] to extract estimated depth maps from the input video. For the 2D style translator, we adopt DCT-Net [30] and VToonify [53] to produce target stylized images and preserve forward/backward facing results whose absolute yaw angle is less than 0.2 radian for supervision. We use the Adam optimizer [22] with the learning rate of 2.5e-5 to train our models and sample 512 rays for each batch. The loss weights are shared by three stages with $\lambda_{mask}, \lambda_{mvs}, \lambda_{reg}, \lambda_{ds}$ set to $\{0.5, 0.5, 0.1, 1\}$. Stage I, II and III are trained for 300k, 200k and 50k iterations, respectively, taking around 20 hours in total on a single NVIDIA Teasla-V100 GPU.

Datasets. We create a 360° captured portrait dataset called Portrait360 to evaluate our approach. This dataset contains 14 static portrait videos captured by rotating the camera around the human head. All videos have a length between 20 to 30 seconds and are split to 300 frames as source training data.

4.1. 3D portrait stylization

Performance on view consistent rendering. Given a short portrait video captured by a monocular camera, our model learns a stylized 3D representation from 2D portrait frames. Stylized portrait images can be generated from arbitrary novel viewpoints following exemplar styles, while ensuring facial identity of the person and 3D consistency between different views. Note that the synthesized images in this part are produced directly by volume rendering on implicit functions, without any explicit style enhancement applied for the results. Our stylized avatars rendered in novel viewpoints and their corresponding exported meshes are shown in Figure 4, more results can be found in the supplementary.

Comparison with 3D avatar stylization methods. In this section, we compare our method with two 3D avatar stylization methods, DeformToon3D [57] and NeRF-Art [48], which represent the state-of-the-art techniques in 3D-aware generative toonification and text-guided NeRF stylization, respectively.

Qualitative comparison. Here we adapt VToonify [53] to generate target stylized images with selected exemplars to train our model. For DeformToon3D [57], we use the author-provided code and train the model using data generated with the same exemplars by DualStyleGAN [52], which is also the 2D generator used in VToonify. Here we directly generate its real-space and style-space results under the same sampled instance code, since the additional PTI [43] process will cause accumulated fidelity errors, especially on arbitrary real faces. For NeRF-Art [48], as it does not support using a single exemplar image for style guidance, we use Mini-GPT4 [59] to generate style de-

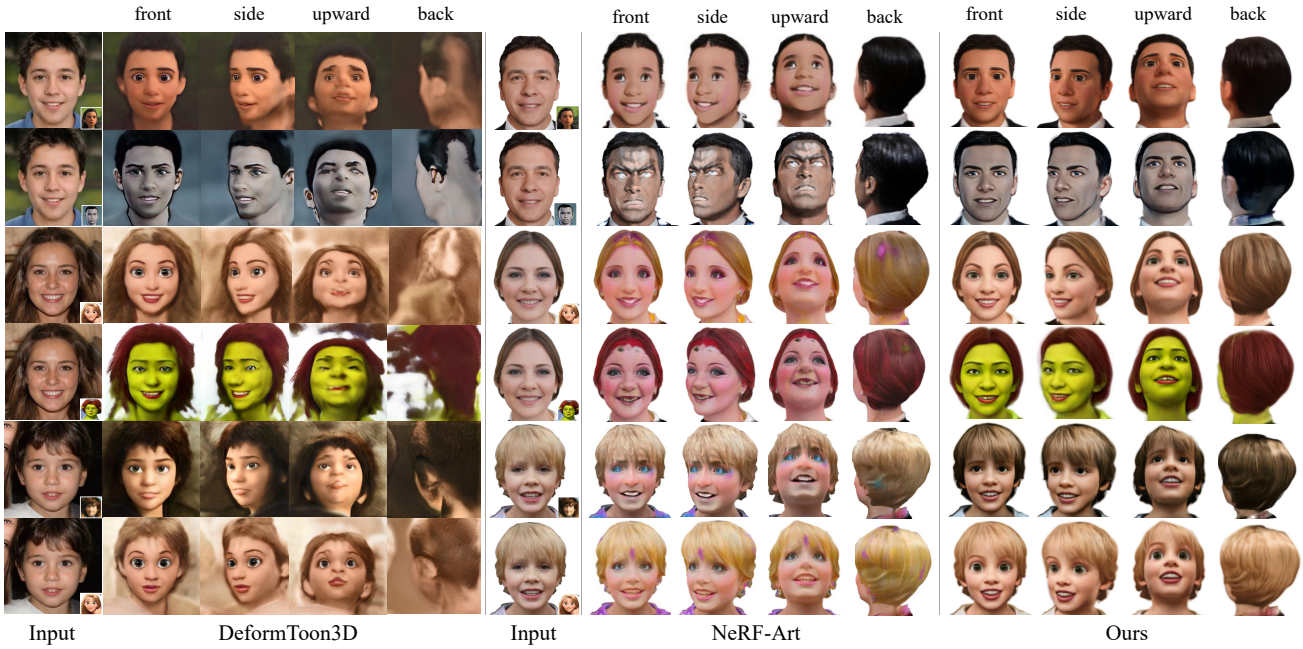


Figure 5. Qualitative comparison with 3D avatar stylization methods. We directly compare the generated real-space and style-space results of DeformToon3D to alleviate the fidelity loss in the additional PTI process. The models of NeRF-Art and Ours are trained on our Portrait360 dataset. Four views are selected for comparison.

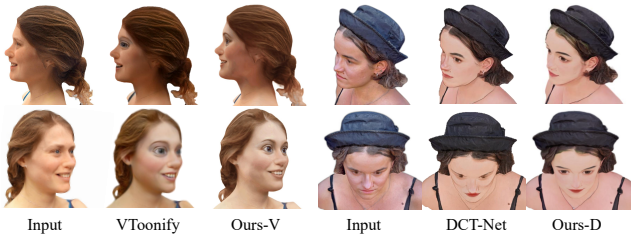


Figure 6. Qualitative comparison with 2D portrait stylization methods on view consistent rendering. For a more prominent video comparison, please refer to the supplementary video.

scriptions corresponding to each target image. The input text prompts used in this section are shown in the supplementary. We demonstrate qualitative comparison of the three methods in Figure 5. DeformToon3D only focuses on frontal views and fails to generate plausible renderings under large angles. Besides, it tends to synthesize overly exaggerated results and fail to maintain the facial characteristics (e.g., hairstyles) of the original image. NeRF-Art only generates results with undesired stylized texture and weakly-changed underlying structures. On the contrary, our method can generate fine-grained full-head stylized avatars with view-consistent renderings and exaggerated styles.

Quantitative comparison. For quantitative comparison, we measure the quality of multi-view stylized renderings of all methods by calculating the Frechet Inception Distance (FID) [14] value for the training cartoon exemplar dataset. A lower FID score indicates that the distribution of the generated images is more similar to that of real 2D cartoon faces. we also evaluate the fidelity of all methods in 3D style adaption using the identity preservation (IP) metric, which is calculated as the Arcface [9] feature similarity between the input image and the stylized result. As shown in Table 1, our method outperforms the other two methods in both FID and identity preservation, which showcases our ability of generating high-quality stylized results while being faithful to the original human identity.

Table 1. Quantitative comparison with 3D avatar stylization methods on FID and IP. ↑, ↓ denote if higher or lower is better.

Method	DeformToon3D	NeRF-Art	Ours
FID ↓	66.5	78.8	**57.6**
IP ↑	0.551	0.671	**0.678**

Comparison with 2D portrait stylization methods. In this section, we compare our method with two state-of-the-art 2D portrait stylization methods, VToonify [53] and DCT-Net [30], to further demonstrate our ability of generating 3D-consistent and high-quality stylized results for arbitrary views.

Qualitative comparison. Due to the incapability of 2D portrait stylization methods to synthesize novel view results, we only make comparison under reconstructed views captured in the input video. For both VToonify and DCT-Net, frames are directly input into the trained/finetuned models released by authors to obtain the corresponding stylized images. Then we select their forward/backward results as sparse view supervision to train our models (denoted as ours-V and ours-D, respectively). As illustrated in Figure 6, VToonify and DCT-Net fail to synthesize exaggerated ge-

Table 2. Comparison of FID and 3D validity with 2D portrait stylization methods.

Method	DCT-Net	Ours-D	VToonify	Ours-V
FID ↓	126.1	**94.7**	86.9	**57.6**
3D validity ↑	0.54	**1.00**	0.62	**1.00**

Table 3. Ablation of the progressive training scheme. Results verify the effectiveness of the proposed module in each stage.

Variants	w/o Prior	w/o GA	w/o TA	w/o PSA	full model
FID ↓	98.7	105.2	96.7	96.2	**94.7**

ometry effects in challenging viewpoints (e.g., side faces) and are unable to maintain 3D view consistency. Note that these extreme view results are not used as supervision in our style adaption process. On the contrary, our method can easily render style-faithful and robust results in a 3D consistent manner. This showcases the importance of learning underlying 3D structures in maintaining view-consistency of the stylized avatar.

Quantitative comparison. We also measure the quality of our rendering results against VToonify [53] and DCT-Net [30] using FID [14]. We use data from our Portrait360 dataset as source images and remove failure cases of the 2D methods. As shown in Table 2, both of our models produce better results with lower FID values compared with original 2D methods. To further evaluate the stylization ability of handling views from the entire 3D space, we propose to calculate 3D validity by computing the conversion rate of successfully stylized results to the whole dataset. 2D methods rely on detected facial landmarks and failed conversions can be automatically recognized. Compared to 2D methods, our method could handle more challenging poses in the entire 3D space with higher 3D validity.

4.2. Ablation study

In addition to visualized results in Figure 3, we verify the effectiveness of the proposed module in each stage by evaluating the performance of corresponding variants of our method. The qualitative and quantitative results are shown in Figure 7 and Table 3, respectively.

MVS guided prior learning. We train a model without photorealistic prior learning and directly learn the spatial neural representation from stylized portrait images. It is confusing for inverse rendering to produce valid geometry and texture with unreal 3D-inconsistent stylized observations, as shown in Figure 7 (b). This indicates that the reconstruction prior is crucial for generating plausible underlying structures in 3D style adaption. The design of MVS guidance also helps to reconstruct more robust surface without holes brought by illumination noise in complicated real-world scenes (see Figure 7 (g)).

Progressive structure adaption (PSA). By removing PSA proposed in Section 3.3, we jointly learn the geometry and texture adaption with the full SNR network. Results in

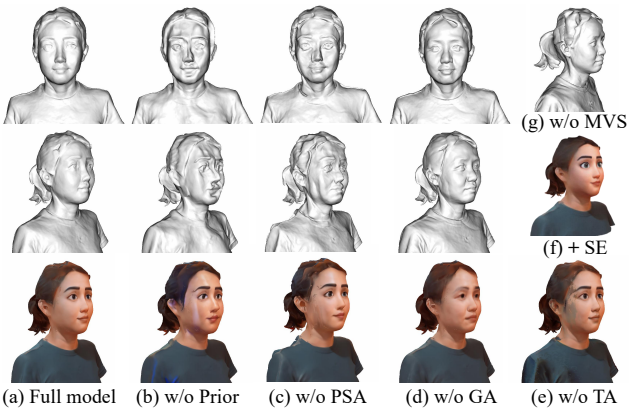


Figure 7. Effects of the proposed prior learning (Prior), progressive structure adaption (PSA), geometry adaption (GA), texture adaption (TA), style enhancement (SE) and MVS guidance (MVS).

Figure 7 (c) show that simultaneously training F_{geo} and F_{tex} disrupts the disentanglement of each other. Progressive adaption brings more accurate surfaces and seamless textures.

Geometry and texture adaption. Figure 7 (d, e) verify the necessity of geometry adaption (GA) and texture adaption (TA), respectively. In contrast to explicit texture stylization, GA enables the internal surface to be deformed adaptively, thus making 3D portraits be fully stylized. Without TA, inferred vertex colors from the appearance field suffers noticeable artifacts, due to the inconsistent observed colors from different views. TA introduces an extra texture field that automatically decomposes albedo colors shared in 3D space, thus alleviating the texture seaming issue. Besides, we explore adding additional style enhancement (SE) on the explicit texture map extracted from the texture field, which further brings more vivid stylization effects (Figure 7 (f)). We also show the impact of the number of stylized frames used for adaption stages in the supplemental material.

5. Conclusion

In this paper, we handled the challenging and on-going task of synthesizing the high-fidelity stylized 3D avatar from a portrait video under the guidance of a single style image. We showed that the naïve combination of portrait style transfer and 3D reconstruction techniques does not work well in this task, and proposed a novel framework called 3DToonify that learns 3D style adaption based on spatial neural representations (SNR). We introduced a delicately-designed spatial neural network for disentangled geometry and texture adaption. We also came up with a novel progressive training scheme suitable for the SNR to accurately capture the underlying stylized 3D structures. Both qualitative and quantitative experimental results demonstrated that our method enables fine-grained 3D avatar stylization with view consistency and diverse exaggerated results.

References

[1] Openmvs. [EB/OL]. https://github.com/cdcseacave/openMVS/. 6

[2] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3davatargan: Bridging domains for personalized editable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4552–4562, 2023. 2, 3

[3] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2565–2574, 2020. 6

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2, 3

[5] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 3

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2, 3

[7] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 2

[8] Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. Stylizing 3d scene via implicit representation and hypernetwork. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1475–1484, 2022. 2

[9] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 7

[10] 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)*, 41(4):1–13, 2022. 2

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[12] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 3, 5

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

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7, 8

[15] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *arXiv preprint arXiv:2305.12529*, 2023. 2

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2

[19] Gwanghyun Kim and Se Young Chun. Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14203–14213, 2023. 3

[20] Gwanghyun Kim, Ji Ha Jang, and Se Young Chun. Podia-3d: Domain adaptation of 3d generative model across large domain gap using pose-preserved text-to-image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22603–22612, 2023. 3

[21] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations*, 2020. 2

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *arXiv preprint arXiv:2306.09329*, 2023. 2

[24] Bing Li, Yuanlue Zhu, Yitong Wang, Chia-Wen Lin, Bernard Ghanem, and Linlin Shen. Anigan: Style-guided generative adversarial networks for unsupervised anime face generation. *IEEE Transactions on Multimedia*, 2021. 2

[25] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 2

[26] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5871–5880, 2020. 2

[27] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 2

[28] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8572, 2020. 3

[29] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 3

[30] Yifang Men, Yuan Yao, Miaomiao Cui, Zhouhui Lian, and Xuansong Xie. Dct-net: domain-calibrated translation for portrait stylization. *ACM Transactions on Graphics (TOG)*, 41(4):1–9, 2022. 1, 2, 5, 6, 7, 8

[31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[32] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3, 4, 6

[33] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2, 3

[34] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *arXiv preprint arXiv:2207.02363*, 2022. 2

[35] Thu Nguyen-Phuoc, Gabriel Schwartz, Yuting Ye, Stephen Lombardi, and Lei Xiao. Alteredavatar: Stylizing dynamic 3d avatars with fast style adaptation. *arXiv preprint arXiv:2305.19245*, 2023. 2, 3

[36] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in Neural Information Processing Systems*, 33:6767–6778, 2020. 2

[37] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 3

[38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 2, 3

[39] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3

[40] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2, 3

[41] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[42] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[43] 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. 6

[44] 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

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[47] Shen Sang, Tiancheng Zhi, Guoxian Song, Minghao Liu, Chunpong Lai, Jing Liu, Xiang Wen, James Davis, and Linjie Luo. Agileavatar: Stylized 3d avatar creation via cascaded domain bridging. 2022. 3

[48] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2, 3, 6

[49] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2, 3, 4, 6

[50] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2

[51] Shiyao Xu, Lingzhi Li, Li Shen, Yifang Men, and Zhouhui Lian. Your3demoji: Creating personalized emojis via one-shot 3d-aware cartoon avatar synthesis. In *SIGGRAPH Asia 2022 Technical Communications*, pages 1–4. 2022. 2

[52] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2022. 6

[53] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Vtoonify: Controllable high-resolution portrait video style transfer. *arXiv preprint arXiv:2209.11224*, 2022. 1, 2, 5, 6, 7, 8

[54] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 3

[55] Chi Zhang, Yiwen Chen, Yijun Fu, Zhenglin Zhou, Gang Yu, Billzb Wang, Bin Fu, Tao Chen, Guosheng Lin, and Chunhua Shen. Styleavatar3d: Leveraging image-text diffusion models for high-fidelity 3d avatar generation. *arXiv preprint arXiv:2305.19012*, 2023. 2

[56] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose. *arXiv preprint arXiv:2308.03610*, 2023. 2

[57] Junzhe Zhang, Yushi Lan, Shuai Yang, Fangzhou Hong, Quan Wang, Chai Kiat Yeo, Ziwei Liu, and Chen Change Loy. Deformtoon3d: Deformable neural radiance fields for 3d toonification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9144–9154, 2023. 2, 3, 6

[58] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *European Conference on Computer Vision*, pages 717–733. Springer, 2022. 2

[59] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 6

[60] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2