HyperStyle3D: Text-Guided 3D Portrait Stylization via Hypernetworks

Zhuo Chen, Xudong Xu, Yichao Yan, Zhengqin Xu, Ye Pan, Wenhan Zhu,
Wayne Wu, Bo Dai, Xiaokang Yang

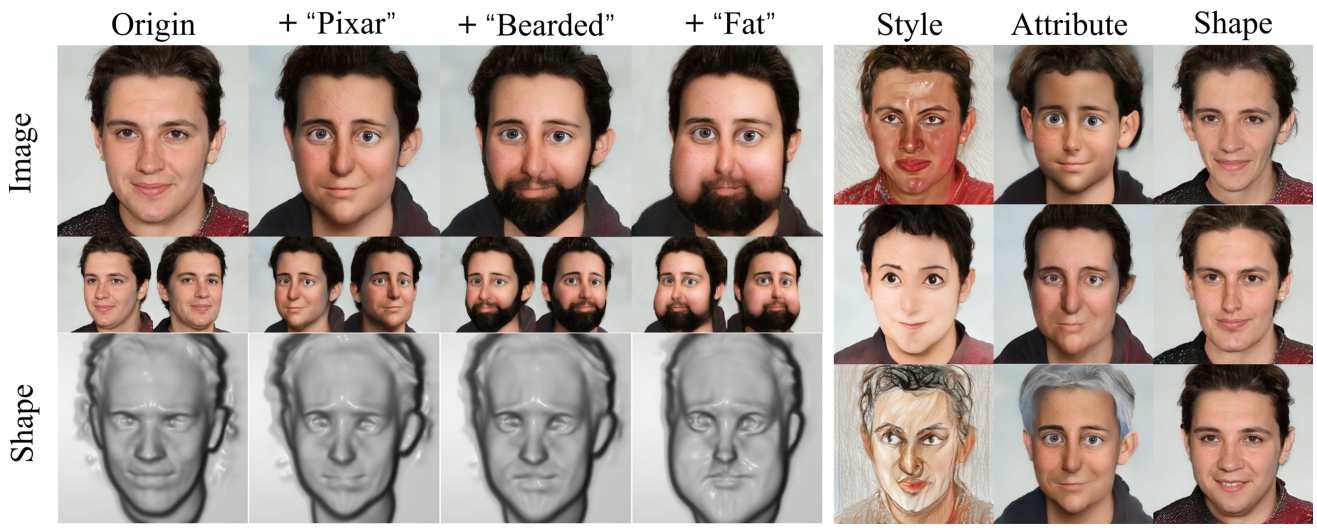


Fig. 1: Examples of text-guided **3D portrait stylization**. Our model enables text-guided style transfer, attribute editing, shape deformation, and their overlying manipulations. Our project page: https://windlikestone.github.io/HyperStyle3D-website/.

***Abstract*—Portrait stylization is a long-standing task enabling extensive applications. Although 2D-based methods have made great progress in recent years, real-world applications such as metaverse and games often demand 3D content. On the other hand, the requirement of 3D data, which is costly to acquire, significantly impedes the development of 3D portrait stylization methods. In this paper, inspired by the success of 3D-aware GANs that bridge 2D and 3D domains with 3D fields as the intermediate representation for rendering 2D images, we propose a novel method, dubbed HyperStyle3D, based on 3D-aware GANs for 3D portrait stylization. At the core of our method is a hyper-network learned to manipulate the parameters of the generator in a single forward pass. It not only offers a strong capacity to handle multiple styles with a single model, but also enables flexible fine-grained stylization that affects only texture, shape, or local part of the portrait. While the use of 3D-aware GANs bypasses the requirement of 3D data, we further alleviate the necessity of style images with the CLIP model being the style guidance. We conduct an extensive set of experiments across the style, attribute, and shape, and meanwhile, measure the 3D consistency. These experiments demonstrate the superior capability of our HyperStyle3D model in rendering 3D-consistent images in diverse styles, deforming the face shape, and editing various attributes.**

(Corresponding author: Yichao Yan.)

Zhuo Chen, Yichao Yan, Zhengqin Xu, Ye Pan, Wenhan Zhu and Xiaokang Yang are with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China. (email: ningci5252@sjtu.edu.cn, yanyichao@sjtu.edu.cn, zhengqinxu@sjtu.edu.cn, whitneypanye@sjtu.edu.cn, zhuwenhan823@sjtu.edu.cn, xkyang@sjtu.edu.cn)

Xudong Xu, Wayne Wu and Bo Dai are with the Shanghai AI Laboratory, Shanghai, China.

***Index Terms*—3D-aware GAN, Style Transfer, Hyper-network.**

I. Introduction

Portrait stylization, pervasive in the entertainment industry, aims at transferring the photorealistic face into a target style while preserving the original identity. Automatic strategies of neural style transfer, which transfer the given artistic style to a photo based on deep neural networks, have demonstrated the capability of bridging the texture gap between two domains. The impressive effect of style transfer gives rise to a wide range of potential prospects, *i.e.*, , virtual makeup, 3D cartoon creation, and short-form video.

Comprehensively considering the application scenarios, an ideal method for portrait stylization is supposed to satisfy three key orthogonal characters, (i) **natural appearance on style**, (ii) **flexible deformation on 3D shape**, and (iii) **light reliance on data.** In other words, high-quality appearance is the fundamental requirement of style transfer, while the ability of shape deformation contributes to a wider range of applications related to 3D creation, and the light reliance on data can lower the threshold for model training.

Currently, 2D-based methods [8], [9], [15], [20], [31], [32], [43], [56], [70] are capable of rendering the image in the target style while preserving the source content, but portrait stylization is more than just matching the texture style. Despite

the remarkable appearance after stylization, these models fall short of shape deformation and multi-view consistency. 3D stylization [11], [21], [28] directly operates on the mesh or the voxel, showing the ability to deform shape and compelling 3D consistency on the rendered multi-view images. However, these methods need dense 3D data of the target style, which demands plenty of effort from professional artists, making such data extremely rare and expensive. Moreover, these 3D methods are single-style, limited by the specific network and 3D data, and thus incapable of adapting to diverse styles in real-world applications.

Inspired by the recent attempts at stylization in 3D-aware GANs [55], [68], we propose to capitalize on the generative radiance fields for zero-shot portrait stylization, by predicting the parameter offsets of generator driven by the text prompt. Based on the powerful implicit 3D representations like NeRF [41], [63], [66] and tri-planes [2], these 3D-aware GAN models could synthesize high-resolution and 3D consistent images by learning from unposed single-view 2D images only and simultaneously infer the corresponding high-quality 3D shapes, providing a potential model for us to achieve all the three characters of ideal portrait stylization. With a similar training pipeline as 2D GANs, 3D-aware GANs can leverage the mature style transfer techniques in the 2D domain for appearance. More importantly, these methods are also empowered to edit the underlying 3D shapes while bypassing the demands of 3D data. Meanwhile, in the 2D domain, CLIP-based [45] methods open a new venue for the text-driven manipulation of StyleGAN [8], [43], and thereby enable the 3D-aware model to realize zero-shot stylization. However, employing 2D methods on 3D-aware GANs still faces several challenges. The optimization of generator parameters [8] or the utilization of latent mapper [43] for a particular input text prompt leads to (i) **single-style limitation** and (ii) **incapacity of overlying manipulations**.

To further address the challenges above, we propose HyperStyle3D, an efficient architecture that leverages the hyper-network to bridge the CLIP model and 3D-aware GANs for high-quality zero-shot portrait stylization. Our hyper-network directly predicts the offset of parameters, getting rid of the time-consuming optimization phase. The learning capacity of the hyper-network enables it to embed diverse manipulations simultaneously, and hence we integrate multiple styles in a unified hyper-networks, solving the problem of weak generalization confronted by such single-style methods. Furthermore, the flexibility of the hyper-network gives rise to another benefit we can explicitly separate different layers of the hyper-network and the generator into multiple levels. We study the properties of the facial semantics learned by different layers of the generator and find that face shape, attributes, and appearance style are orthogonally controlled by three different levels of the generator. Therefore, we split the layers into three groups, *i.e.*, , coarse, medium, and fine, whose parameters offset are predicted by three corresponding layer groups of our hyper-network, responsible to shape deformation, attribute editing, and general style transfer, respectively. Based on this, our hyper-network-based framework bears an additional advantage of overlying manipulations driven by texts with multi-level semantics.

Thanks to the proposed hyper-network, our zero-shot 3D-aware stylization model, dubbed HyperStyle3D, overcomes the challenges during the utilization of 3D-aware GANs and successfully meets all three aforementioned characters of an ideal portrait stylization. In summary, the contribution of this paper is threefold:

- We introduce the CLIP model to 3D-aware GANs, realizing the zero-shot 3D portrait stylization while avoiding the dependence on expensive 3D data.
- A novel hyper-network is proposed for multi-style portrait stylization, which enables style transfer, attribute manipulation, shape deformation, and their overlying manipulations in a unified framework.
- We demonstrate high-quality results on 3D portrait stylization in terms of appearance, shape and 3D consistency as shown in Fig. 1.

II. Related Work

2D Style Transfer. 2D style transfer is an image processing method that renders the semantic content of the image with different styles. A wide variety of image transformation tasks [9], [15], [20], [31]–[34], [56], [59] has been proposed based on Convolution Neural Networks (CNN). Gatys *et al.* [9] first introduces the CNN to the task of style transfer and successfully splits the features into content and style by structuring the GRAM matrix. For fast inference during arbitrary style transfer, AdaIN [15] performs style transformations by switching means and variances of the feature. With the popularity of 2D GANs, the task of style transfer gradually spreads to unsupervised generative models. CycleGAN [70] design a pair of GANs, capable to use the unpaired training data for domain adaptations. With the advent of StyleGAN [22]–[24], plentiful works [8], [13], [14], [18], [29], [30], [40], [43], [44], [47], [67] propose to synthesize high-quality style images based on this hierarchical style-adaptive framework. Recently, to relieve the requirement of style images as training data, StyleCLIP [43] proposes to use a text to discover global directions in the latent space. Rather than exploring the latent space, StyleGAN-Nada [8] fine-tunes the generator for large changes in the style, beyond the generator's original domain. In this work, we extend the style transfer of 2D GANs to 3D-aware GANs, enabling 3D shape deformation and great 3D consistency.

Latent Manipulation. Many works explore the latent space of a pre-trained generator for image manipulation [1], [19], [51]–[53], [57], [71]. These approaches can be roughly classified into two categories, i.e., 1) unsupervised methods that explore the semantics of generator to discover distinguishable directions [12], [53], [57] and 2) Supervised methods that use attribute labels to find meaningful latent path [1], [51], [52], [71]. Sefa [53] proposes a closed-form factorization algorithm to discover latent semantic directions by directly decomposing the pre-trained weights. StyleFlow [1] utilizes conditional normalizing flow to disentangle the attributes and broaden the editable attributes. Overall, the manipulation in latent space has demonstrated the ability to perform precise

editing of specific attributes. However, it's not easy to discover the disentangled path in latent space for arbitrary manipulation. Our HyperStyle3D utilizes hyper-networks to explore the editability in the parameter space of the generator for overlying manipulation of arbitrary styles, attributes, and shapes.

3D Stylization. Editing 3D content according to a given style is a challenging task [21], [61], [64], [69], involving both geometric deformation and texture transformation. 3DStyleNet [64] achieves both shape deformation and texture stylization by a part-aware affine transformation field for shape and a multi-view differentiable renderer for texture editing. Yang *et al*. [21] utilizes detailed 3D mesh data and 2D image data of caricature style to train a deformable 3D caricature framework. With the advent of NeRF [35] and the CLIP model [45], researchers have already made remarkable progress on 3D text-driven synthesis [16], [17], [36], [49], [58]. These methods adapt the optimization procedures supervised by the CLIP model [45]. Specifically, CLIP-NeRF [58] proposes a unified framework to manipulate NeRF, guided by a text prompt or an example image. Given a 3D mesh, Text2Mesh [36] can modify the color and geometry under the guidance of a text prompt. However, these text-driven methods are all per-instance stylizations that cannot generalize to all objects belonging to a category. In contrast, our HyperStyle3D is a generative model that edits the face style via learned parameters, capable to apply the manipulation to all the faces. Recently, Diffusion models are introduced as a style prior to 3D stylization. DATID-3D [25] and PODIA-3D [26] both leverages diffusion models to generate 2D style images as the training data to fine-tune the pre-trained 3D GANs. As the synthetic training data are arbitrary and hard to control quality, these methods additionally propose data filters to eliminate the low-quality data. However, despite diverse generation, these methods always fail to maintain the identity feature during the domain adaptation. Compared to them, our methods can achieve both style transfer and identity consistency.

3D-Aware Image Synthesis. Inspired by the superiority of NeRF [37] representation, Several attempts [2], [3], [6], [7], [10], [27], [38], [39], [41], [42], [46], [50], [54], [60], [62], [63], [68] deploy radiance fields into generative models and thus enable 3D consistent image synthesis. Recently, 3D-aware GANs [2], [10], [41], [60], [63], [68] combine the shallow NeRF features with a 2D CNN-based renderer and could hallucinate images at 1024^2 resolution. In particular, StyleSDF [41] takes Signed Distance Fields (SDF) as the intermediate representation for high-quality shape generation and introduces the regularization loss to preserve 3D consistency. Nonetheless, expressive 3D-aware GANs have a huge demand for image data and computational sources, arduous to re-train a generator according to the desirable styles. In our work, we leverage a text-driven zero-shot method to adapt the 3D-aware GANs across various domains, reducing the dependency on style data.

III. Method

Our work aims to enable a pre-trained 3D-aware generator with multi-level manipulations including style transfer, attribute editing, and shape deformation. For a specific manipulation, our zero-shot method leverages a hyper-network to predict the parameter offsets of the pre-trained generator, under the guidance of corresponding text prompts. For multi-level manipulations, we split the layers of our hyper-network into multiple groups, each of which corresponds to a specific-level manipulation. The overview of our architecture is shown in Fig. 2.

A. Preliminaries on Generative Radiance Fields

Adopting NeRF [37] as the 3D representation for 3D-aware GANs has been explored in several works [2], [3], [6], [7], [10], [27], [38], [39], [41], [42], [46], [50], [54], [60], [62], [63], [68]. Here we follow the design of StyleSDF [41], one of these awesome models, as the pre-trained generator of the source domain. StyleSDF is one of the typical NeRF-based GAN models, which adopts the implicit function parameterized as an MLP to represent the 3D scene. This function takes a 3D coordinate $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$ and the view direction $\mathbf{d} \in \mathbb{S}^2$ as inputs and outputs per-point signed distance values $\alpha(x) \in \mathbb{R}^+$ and view-dependent color $\mathbf{c}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^3$.

Unlike pure-MLP generator, instead of directly computing each pixel color for image generation, StyleSDF additionally computes a low-resolution feature map $\mathbf{f}(\mathbf{x}, \mathbf{d})$ via volume rendering along its corresponding camera ray. To achieve high-resolution generation under the constraint of memory, it utilizes a similar up-sampling architecture as StyleGAN to efficiently transformed the low-resolution feature maps into high-resolution images. We summarize this image generation process as $\mathbf{I}_g = G_\theta(\mathbf{z}, \boldsymbol{\xi})$, where G_θ is the generator, θ denotes its learnable parameters, $\mathbf{z}$ is the latent code conditioning the generator via mapping networks and $\boldsymbol{\xi}$ represents the sampled camera pose.

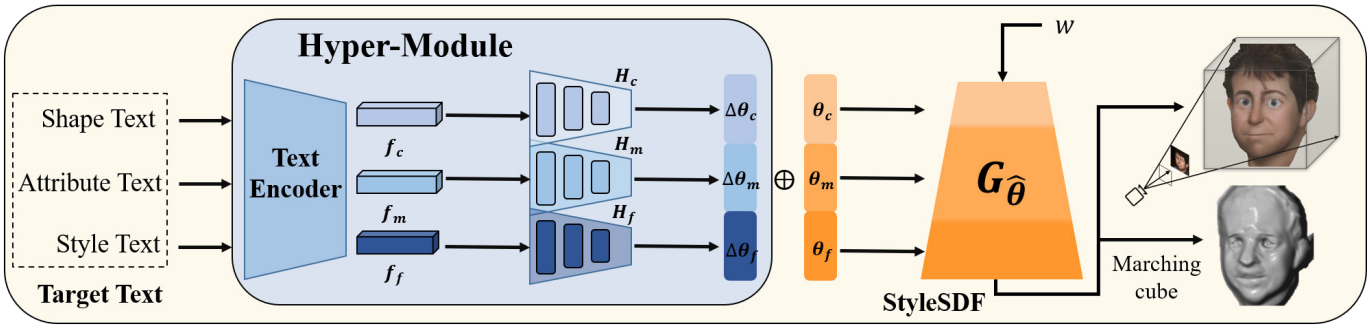
The architecture of StyleSDF is shown in Fig. 3 (a).

B. Single-forward Transfer via Hyper-Module

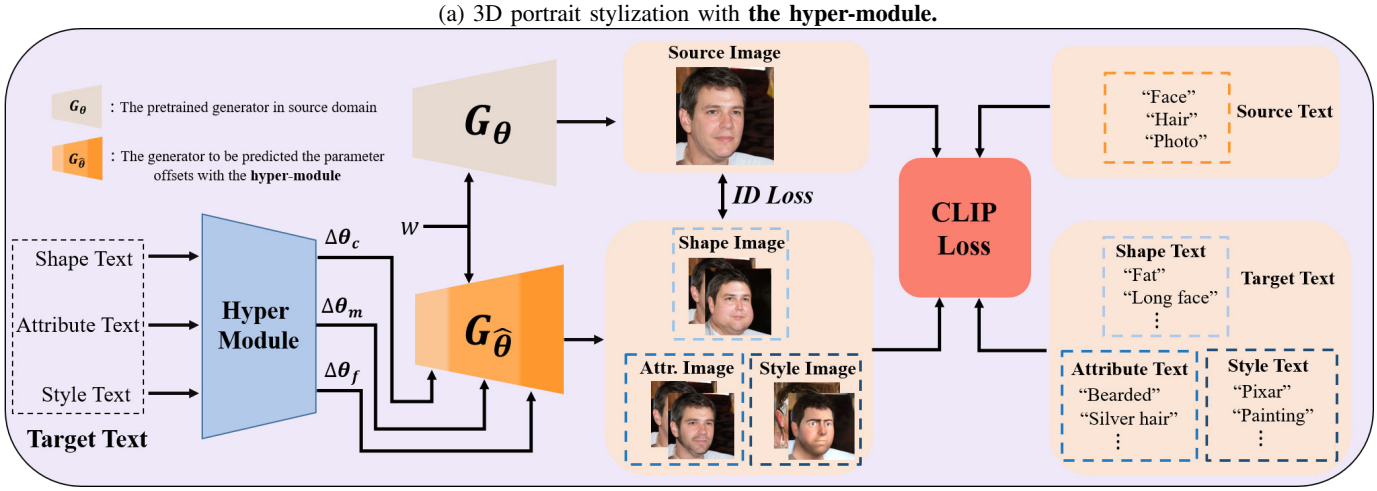
Previous GAN-based face methods usually perform image manipulation in the $\mathcal{W}$ or $\mathcal{W}+$ spaces, however, it is non-trivial to find the specific latent path for the subtle attributions like eyes size without labeled data. It is also difficult to disentangle the latent path, especially for multi-level manipulation. Additionally, as different identities correspond to different w codes, the semantic direction in latent space is usually instance-specific.

In contrast, parameter space contains more essential and disentangled facial semantics learned by the generator. Thus, we explore the editability in the parameter space of GANs to discover a general, detailed and disentangled editing path with the guidance of text prompts. Unlike other methods using per-style optimization process [8], [43] to fine-tune the parameters, we instead design a single-forward framework based on the hyper-network to predict the parameter offsets of the generator G_θ to simultaneously support multiple styles. The refining process is guided by the changing direction of the source text and the target text.

Specifically, we first use a shared text encoder E to transform source text t_{src} and target prompts t_{tgt} into the



(a) 3D portrait stylization with **the hyper-module.**



(b) **Training Pipeline** with multiple styles integrating into a hyper-module.

Fig. 2: **The overview** of our full-pipeline. (a) Our hyper-module consists of three trainable hyper-networks and a fixed text encoder. The text prompts of three levels (shape, attribute, and style) are encoded into the coarse, medium, and fine direction features, which are then fed into the corresponding hyper-network. The hyper-networks predict three groups of parameter offsets for the coarse, medium, and fine layers in the pre-trained 3D-aware generator. (b) Our hyper-module is trained under the supervision of CLIP loss and ID loss. Text prompts of three levels are simultaneously integrated into the training, empowering the hyper-module to handle the overlying manipulation of diverse styles, attributes, and shapes. We pre-define the source text as a description related to the current training target text, such as "Face" to "Bearded face".

intermediate features, and then represent the manipulation direction as the feature difference:

$$f_{\text{dir}} = f_{\text{tgt}} - f_{\text{src}} = E\left(t_{\text{tgt}}\right) - E\left(t_{\text{src}}\right). \tag{1}$$

As shown in Fig. 2, the shape text, attribute text, and style text are three classes of target texts. Source texts are the prompts corresponding to each target text, for example, "hair" (source) to "silver hair" (target).

To embed the direction code f_{dir} to the generated image $\mathbf{I}_g$, we opt to use f_{dir} to update the pretrained generator G_θ. The updating process is bridged via a series of hyper-networks H. The hyper-network H takes the direction code f_{dir} as input and predicts the parameter offsets $\Delta\theta$, which are then multiplied as a coefficient to refine the parameters of the primary generator G_θ. The offset generator $G_{\hat{\theta}}$ with updated parameters $\hat{\theta}$ enables the generation of manipulated images as the target text describes. Here we only consider predicting the parameters for the main linear layers of G_θ, skipping up-sampling layers, as most of the content is completely generated in the main linear layers.

Assume that the pre-trained generator G_θ has N linear layers with parameters $\theta = (\theta_1; \theta_2; ...; \theta_N)$. As pointed out by the theoretical study of previous works related to the StyleGAN, different layers of the generator contribute to different levels of attributes. Therefore, we propose to use each small hyper-network h_j to predict the parameter offsets $\Delta\theta_j$ of each linear layer j. All the small hyper-networks constitute the integrated hyper-network:

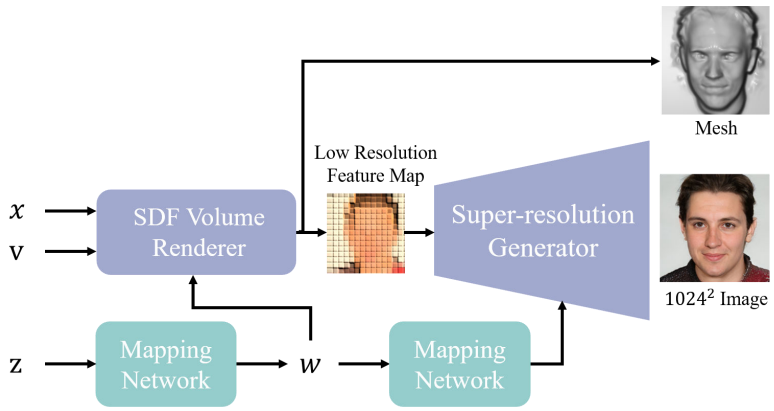
$$H = \{h_j\}, \ j \in \{1, 2, \ldots, N\}. \tag{2}$$

To reduce the risk of over-fitting and enhance the diversity of styles, we additionally add some random noise n^i to each style change direction f^i_{dir}. Thus, the parameter offsets $\Delta\theta_j$ of the linear layer j is predicted as

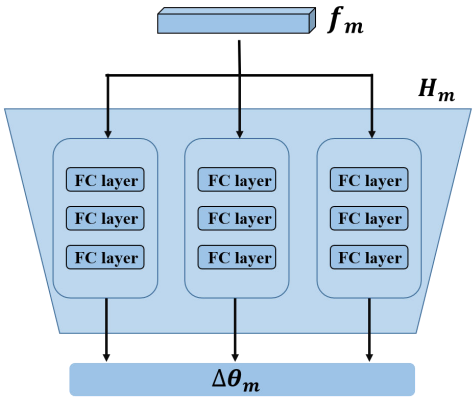
$$\Delta\theta_j = h_j\left(f^i_{\text{dir}} + n^i\right), \ j \in \{1, 2, \ldots, N\}. \tag{3}$$

The details of our hyper-network are illustrated Fig. 2. Finally, the updated generator has the parameter as

$$\hat{\theta} = \{\theta_j \cdot (1 + \Delta\theta_j)\}, \ j \in \{1, 2, \ldots, N\}. \tag{4}$$



(a) The architecture of source generator



(b) The architecture of a group of hyper-network

Fig. 3: Detailed architectures of the networks. **(a)** The architecture of StyleSDF [41] which is composed of a SDF volume renderer and a 2D super-resolution model. **(b)** We take the group of medium hyper-networks as an illustration, and the other two groups of hyper-network also have a similar architecture as the illustration. A group of hyper-networks consists of several small components, each of which is a three-layer MLP.

The manipulated image $\mathbf{I}_g$ could be generated from the updated generator:

$$\mathbf{I}_g = G_{\hat{\theta}}(\mathbf{z}, \boldsymbol{\xi}). \tag{5}$$

As shown in Fig. 3 (b), a group of our hyper-networks consists of several components, each of which is a three-layer light MLP with a hidden layer. We set the channel dimension of the hidden layers to 64. As each layer of the pre-trained generator is predicted by a component, the specific numbers of the component are determined by the layers of pre-trained generator that the group is responsible for.

C. Multi-level Manipulation

As a 3D-lifting architecture based on StyleGAN [22]–[24], StyleSDF [41] displays a similar characteristic that layers make different semantic contributions to the final image generation, and meanwhile this property extends to the 3D geometric contributions due to the 3D representation of the generator. Inspired by the Latent Mapper of StyleCLIP [43], we investigate the effect of different layers and split them into three groups, *i.e.*, , coarse, medium, and fine, according to their contributions, *i.e.*, , shape, attribute, and style, respectively.

Thereby, we extend the hyper-network to also contain three parts along with the three groups of layers. Here, we denote the input text feature as $f = (f_c;\ f_m;\ f_f)$, corresponding to the three groups of layers. The hyper-network is defined by

$$H(f) = \left(H_C\left(f_c\right), H_m\left(f_m\right), H_f\left(f_f\right)\right). \tag{6}$$

Each group of hyper-network responsible to a type of manipulation (shape, attributes, or styles) is trained to predict the parameter offsets, as indicated by the text prompts of corresponding types, while preserving the other visual attributes of the input image.

$$\Delta\theta_c = H_c\left(f_c\right), \Delta\theta_m = H_m\left(f_m\right), \Delta\theta_f = H_f\left(f_f\right). \tag{7}$$

Specifically, as the coarse layers are responsible to the fundamental shape generation, the coarse hyper-network H_c takes several texts related to the geometric deformation such as "fat" as the training target, and learns to search the corresponding path for shape transformation in parameter space by the supervision of CLIP loss. Similarly, the medium layers contribute to the facial topology, and thus the medium hyper-network H_m is trained for attribute editing. Besides, the fine hyper-network H_f is trained for general style transfer because the fine layers mainly influence the color appearance of image results.

The updated parameters are given by:

$$\hat{\theta} = \left(\theta_c + \alpha_c \Delta\theta_c, \theta_m + \alpha_m \Delta\theta_m, \theta_f + \alpha_f \Delta\theta_f\right), \tag{8}$$

where α_* are coefficients indicating the editing degree. As α increases, the synthesized images change more positively along the corresponding manipulation. For example, with text "bearded", the resulting images show a thicker beard as α increases. A negative α guides the results to change in the opposite direction of the text description. The degree coefficients provide more flexibility for editing.

D. Training

CLIP Loss. Similar to StyleGAN-Nada [8], our method to predict the parameter offsets of a generator is mainly guided by a text-image directional objective. The direction loss is given by:

$$\Delta T = E_T\left(t_{\text{tgt}}\right) - E_T\left(t_{\text{src}}\right), \tag{9}$$

$$\Delta I = E_I\left(G_{\hat{\theta}}(w)\right) - E_I\left(G_\theta(w)\right), \tag{10}$$

$$L_{\text{dir}} = 1 - \langle \Delta T,\ \Delta I \rangle. \tag{11}$$

where E_I and E_T are the image and text encoders of the CLIP model, t_{src} is the source class text, t_{tgt} is the input text that indicates the target extrinsic style or intrinsic attribution, and $\langle \cdot \rangle$ is the function of cosine similarity. The idea is to guide

Fig. 4: **Qualitative results**. Our hyper-network-based model can transfer the image into diverse style domains.

the image produced by the updated generator to change only along the indicated text direction.

ID Loss. For intrinsic attributes manipulation, we further leverage an ID loss [5] to preserve the facial identity when predicting the parameter offsets of a generator along the text direction. Besides, we utilize the generated multi-view images to better keep face identity in 3D space. The ID loss L_{ID} is given by:

$$L_{\mathrm{ID}} = \frac{1}{N}\sum_{i}^{N}\left[1 - \left\langle F\left(G_{\hat{\theta}}\left(w, \xi_i\right)\right),\ F\left(G_{\theta}\left(w, \xi_i\right)\right)\right\rangle\right], \tag{12}$$

where $F(\cdot)$ is a pre-trained ArcFace [5] model for face recognition to extract identity features, and i indicates the i-th view direction of the total N views. ID loss encourages the extracted features of source and target images to be as close as possible during editing.

Region Loss. Although the text description only contains the attributions we want to edit, there exist risks that other irrelevant attributions may change as the generator's parameters shift. To address this, we additionally introduce an optional region loss L_{region} to reduce the risk of entangled attributes in parameters space:

$$L_{\mathrm{region}} = \left\| R\left(G_{\hat{\theta}}\left(w, \xi_i\right)\right) - R\left(G_{\theta}\left(w, \xi_i\right)\right)\right\|_2, \tag{13}$$

where $R(\cdot)$ expresses the pixels of irrelevant regions predicted by a pre-trained semantic model BiSenet V2 [65]. The CLIP model is responsible to match the input text to the region. When training the hyper-network for editing attributes, the pixels of irrelevant parts are supervised by the region loss to keep them unchanged as much as possible.

To summarize, the total loss is given by:

$$L_{total} = \lambda_{\mathrm{dir}} L_{\mathrm{dir}} + \lambda_{\mathrm{ID}} L_{\mathrm{ID}} + \lambda_{\mathrm{region}} L_{\mathrm{region}}, \tag{14}$$

where $\lambda_{\mathrm{direction}}$, λ_{ID} and $\lambda_{\mathrm{region}}$ are the weights for CLIP loss, ID loss, and region loss, respectively.

IV. Experiments

A. Training Details

We train our HyperStyle3D model on 1 NVIDIA 3090Ti GPU with a batch size of 4. we use an Adam Optimizer and set the learning rate to 0.0002. Training a group of hyper-networks usually costs 3000 iterations for 6 target styles, roughly 18 minutes. The training phase for an integrated hyper-network is comparable to the optimization procedure [8] in terms of average time per style.

In the training phase, we adopt the iterative training strategy that the hyper-network is alternatively trained for 10 iterations for each text pair. Since three groups of hyper-networks for different levels are orthometric, we train them separately. We find that some target texts usually require longer or shorter iterations than other target texts to achieve expected changes. The requirement of different converging times leads to an unbalanced effect with the unified training time. Therefore, we adopt an adaptive training phase for each style. We set an absolute loss threshold and a relative decreasing threshold to indicate the objective. If the training loss in 10 iterations is invariably less than the absolute threshold and the decreasing degree is less than the relative threshold, we will decline the sampling frequency of this text later in the training.

B. Qualitative results

In this section, we discuss the experiments conducted to evaluate the quality of stylized images.

FFHQ dataset. Our generator starts from the real-image domain of the FFHQ dataset [23], to the multiple out-of-domain styles, *i.e.*, , "Painting", "Elf", "Disney Princess", "Ukiyo-e", "Pixar", "Renaissance" and "Botero". As shown in Fig. 4, our hyper-network-based model can synthesize high-quality stylized images. Although all the styles are embedded in a unified hyper-network, it can be noticed that our model still adapts to a wide range of styles beyond the pre-trained generator's domain. Moreover, to demonstrate the adventages of our method, we compare our HyperStyle3D with the state-of-the-art method DATID-3D and a baseline method that adopts 2D style transfer techniques [48] on the images and then inverts stylized images to a pre-trained 3D-aware GAN [2]. We select two styles whose pre-trained models are provided in DATID-3D, *i.e.*, , Elf and Pixar. As shown in Fig. 5, our HyperStyle3D can synthesis text-consistent styles and natural details while maintaining identity consistency with input faces. Despite style results consistent to the text, DATID-3D unfortunately loses the identity details and make it hard to recognize the original

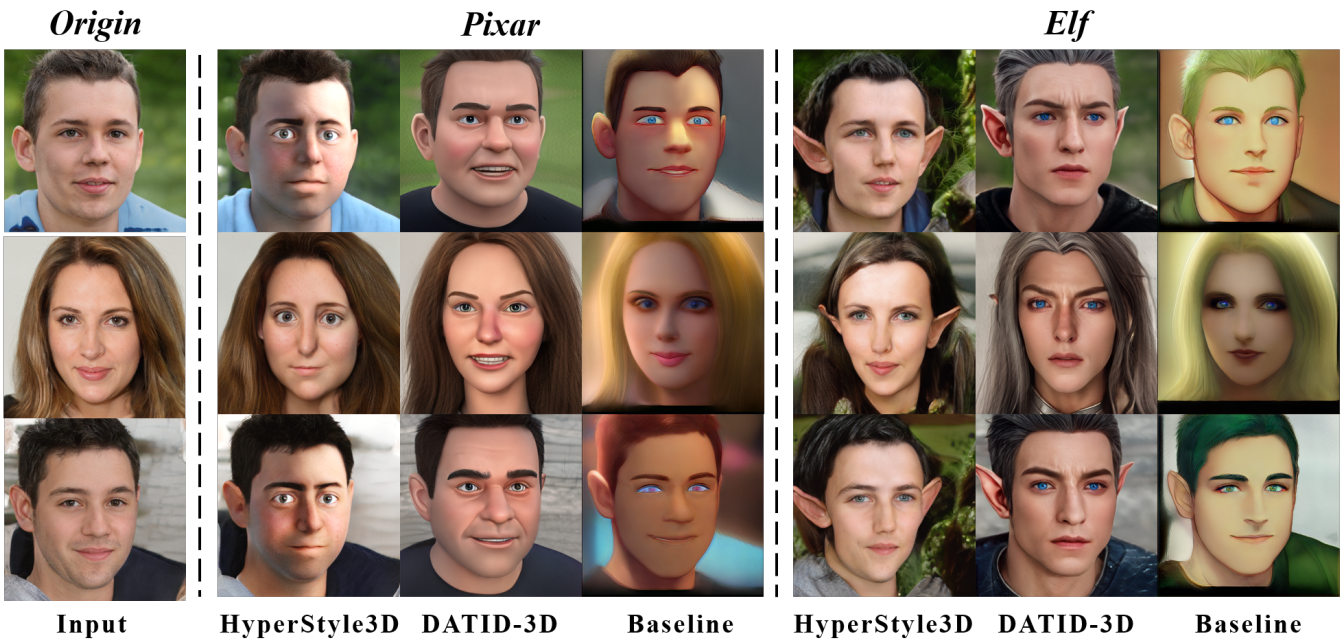


Fig. 5: **Qualitative comparison**. HyperStyle3D and DATID-3D achieves better style realism and recovers more 3D details compared to the baseline, *i.e.*, , 2D transfer + 3D GAN inversion. It's noteworthy that 3D GAN inversion has to operate on the novel style domain, resulting in the loss of rich face details accordingly. Besides, DATID-3D lacks identity consistency during the stylization, while ours maintains better identity feature.

Method	Baseline	DATID-3D	Ours
Style realism ↑	3.6	**4.2**	4.0
Face details ↑	2.8	3.9	**4.1**
ID consistency ↑	3.1	2.9	**3.6**
Training time (9 styles) ↓	**9 × 3 min**	9 × 10 h	30 min
Model size (9 styles) ↓	9 × 377 MB	9 × 377 MB	**512 MB**
Style ID consistency ↑	0.90	0.87	**0.94**

TABLE I: **Quantitative comparisons among baseline, DATID-3D and our HyperStyle3D.** We first conduct a user study to provide an intuitive evaluation of generated style images. Then, we measure the training time and model size to evaluate the efficiency of style transfer. Finally, we measure the similarity to show the identity consistency between origin images and stylized images.

Method	Depth error ↓	ID similarity ↑
Original StyleSDF [41]	1.48	0.81
Hyper-networks for shape	1.52	0.84
Shape + attribute	1.52	0.88
Shape + attribute + style	1.55	0.78

TABLE II: **Quantitative results of 3D consistency.** Our model is based on StyleSDF and shows comparable performance after the stylization. Even with overlying manipulations, our model will not be degraded in terms of depth consistency and ID consistency between various views.

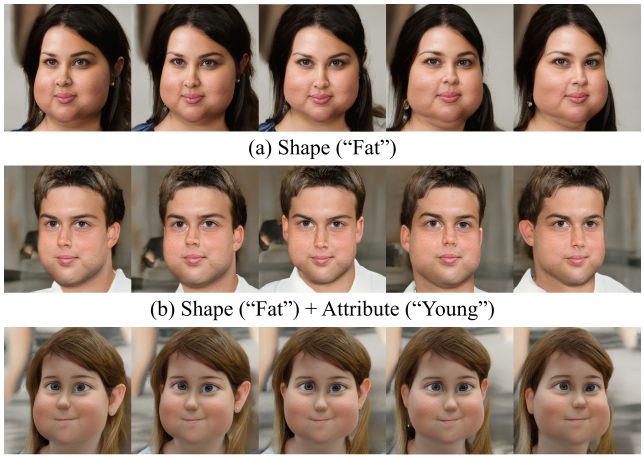
identity. The baseline method can also stylize the portrait under the text guidance, while its results seem too flat and lack a sense of depth. This is because the 2D style transfer ignores the depth dimension, although the 3D GAN inversion transfers it into 3D representation.

C. Quantitative results

User study. To measure the quality of style transfer, we first conduct a user study. Participants were asked to rate the samples on a scale of 1 to 5 based on three criteria, *i.e.*, *(1) Style realism*, *(2) Face details* and *(3) ID consistency*. Style realism is to evaluate the correspondence between generated results and input style prompts. Face detail is to evaluate the appearance quality of results. ID consistency is to evaluate the identity similarity between origin images and stylized images. As shown in Table I, our method is superior to the 2D baseline in both aspects, and achieves comparable rates to DATID-3D. It is noted that our method is better than DATID-3D in ID consistency which is an important target in 3D stylization.

Efficiency. To demonstrate the difference of training hyper-network and fine-tuning generator, we then compare the training time and model size for totally 9 styles. As shown in I, our training is much faster than DATID-3D. Besides, due to the hyper-network, we can learn multiple styles in one model and significantly reduce the model size.

Cross-style Identity Consistency. The target of stylization is to transfer the original portrait into another style while maintaining the identity feature. Therefore, it is necessary to compare the similarity between the portraits before and after the stylization. We adopt the ArcFace [5] cosine similarity as the evaluation metrics in this experiment. As shown in I, simi-



(a) Shape (“Fat”)

(b) Shape (“Fat”) + Attribute (“Young”)

(c) Shape (“Fat”) + Attribute (“Young”) + Style (“Pixar”)

Fig. 6: **The results of 3D consistency**. We sample several view directions to show the 3D consistency. As can be observed, the 3D consistency is still well maintained after the manipulation via hyper-network.

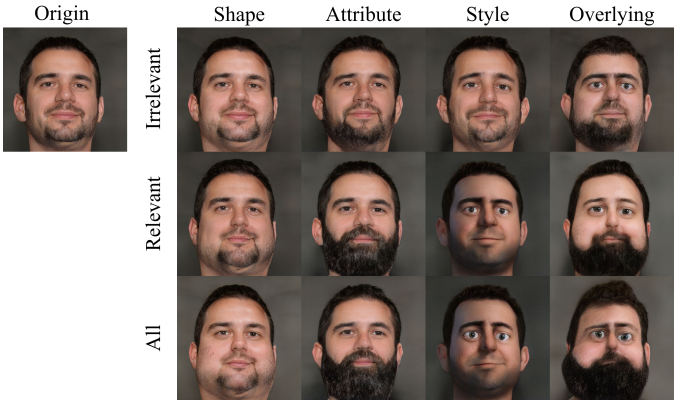


Fig. 7: **Ablation study** of the effect of different layers. For a specific level text prompt, parameter offsets on the relevant (the same level) layers lead to a significant change, otherwise, the change brought by the parameter update on irrelevant layers is less notable. Blindly updating parameters of all layers can work in the single style mode, but results in degenerate images in the overlying manipulation.

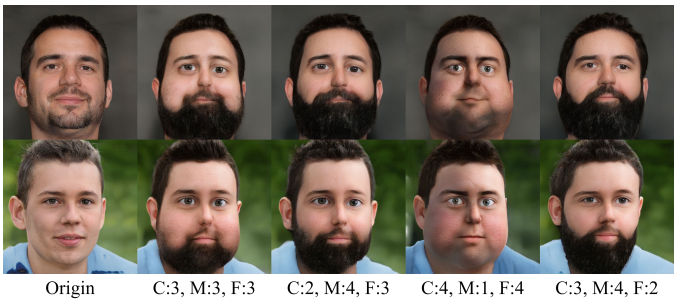


Fig. 8: **Ablation study** on different choices of layer groups. For example, “C:2, M:4, F:3” means that the first two layers serve as the coarse group responsible to shape deformation, the middle four layers are the medium group responsible to attribute manipulation, and the last three layers serve as the fine group responsible to style transfer.

lar to the qualitative result and user study, our HyperStyle3D is better than DATID-3D in cross-style ID consistency measured by the face reorganization model.

3D consistency. One of the advances of 3D stylization compared to the 2D methods is the 3D consistency of generated images. To show that our methods have not undermined the 3D consistency of pre-trained 3D-aware generators, we conduct a quantitative experiment that adopts the depth consistency used in StyleSDF [41] and facial identity consistency used in EG3D [2] as the evaluation metrics of 3D consistency.

Depth Consistency. To evaluate the depth consistency, we sample 500 identities, render their 128×128 depth maps from the frontal view and a random side view, and compute the depth error after the alignment between the two views. A modified Chamfer distance metric is adopted as the measurement of alignment error between two depth points.

$$CD\left(S_1, S_2\right) = \underset{x \in S_1 y \in S_2}{med\min} \|x - y\|_2^2 + \underset{y \in S_2 x \in S_1}{med\min} \|x - y\|_2^2, \tag{15}$$

where S_1 and S_2 are two point clouds of a unified identity from two different views. The metric can better handle the occlusion and background mismatch that is the disturbance during the measuring. As shown in Tab. II, despite stylization and multiple manipulations, our hyper-networks have achieved comparable depth consistency compared to the original StyleSDF [41].

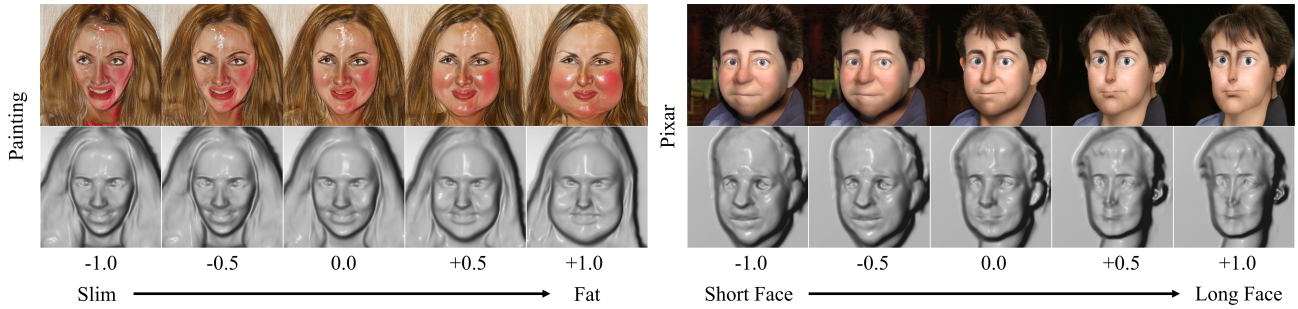
Multi-view Identity Consistency. The ArcFace [5] cosine similarity of facial identity between multi-views is a common metric for the evaluation of 3D consistency adopted in 3D-aware GANs [2], [66].

$$Sim_{ID} = \langle F\left(G_{\hat{\theta}}\left(w, \xi_1\right)\right),\ F\left(G_{\hat{\theta}}\left(w, \xi_2\right)\right)\rangle, \tag{16}$$

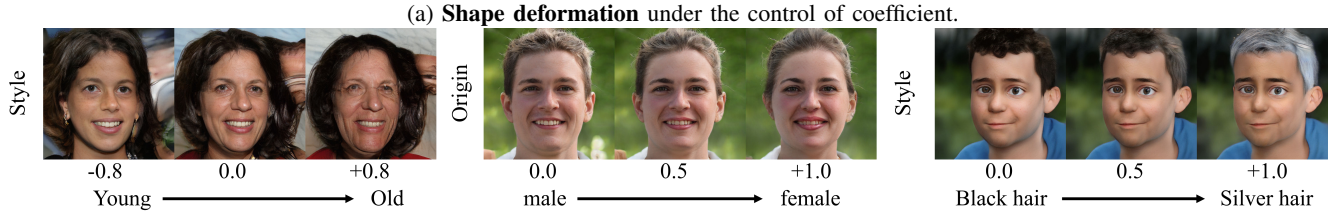
where ξ_1 and ξ_2 are two fixed sampled views, $F\left(\cdot\right)$ is the ArcFace pre-trained model to extract the identity feature, and $\langle\cdot\rangle$ indicates cosine similarity. To evaluate facial identity consistency, we sample 1500 identities, render their 512×512 high-resolution images from two random side views, and measure the cosine similarity of those two views. As shown in Tab. II, after shape deformation and attribute editing, portrait images generated by ours are even superior to those from the original model in terms of ID similarity, demonstrating the preservation of ID consistency of our method. The superior result is because the exaggerated shape and attribute make the facial features more distinctive. We also show several examples with multiple views in Fig. 6 for the illustration of 3D consistency, while additional cases with more views are provided in the *supplementary material*.

D. Ablation Study

Multi-level Manipulation. Due to the different levels of manipulations, it’s beneficial to analyze the layer mechanism and accordingly split hyper-networks into shape-controller, attribute controller, and style-controller. Aiming to achieve disentangled manipulations of multiple levels, we explore the



(a) **Shape deformation** under the control of coefficient.



(b) **Attribute editing** under the control of coefficient.

Fig. 9: **Results of controllable degree of manipulations**. As the coefficient α increases, the manipulated image **gradually** changes towards the target direction. With a negative coefficient, the image varies along the opposite direction of the target text. As we can see, shape and attribute show a similar trend.

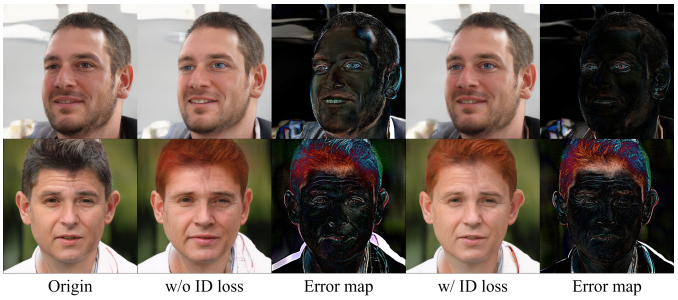


Fig. 10: **Ablation study** on ID loss.

effect of each layer in StyleSDF [41] contributing to image generation. In Fig. 7, we compare the performance of several models with different updated layers according to the different training text prompts. To ensure a fair comparison, all models are trained with the same batch size and learning rate.

As illustrated in Fig. 7, with the text prompt "fat", predicting the parameter offset of the coarse layers leads to significant shape deformation, but updating others is relatively useless. For the text prompts of attribute and style, it also shows a similar trend that fine-tuning medium and fine layers can lead to a significant change in attribute and style, respectively, otherwise, the change brought by the parameter update on irrelevant layers is less notable. Besides, as shown in the last column, with all layers to update, the generator feels confused when each layer simultaneously receives three different parameter offsets predicted by hyper-networks, resulting in erratic generated images.

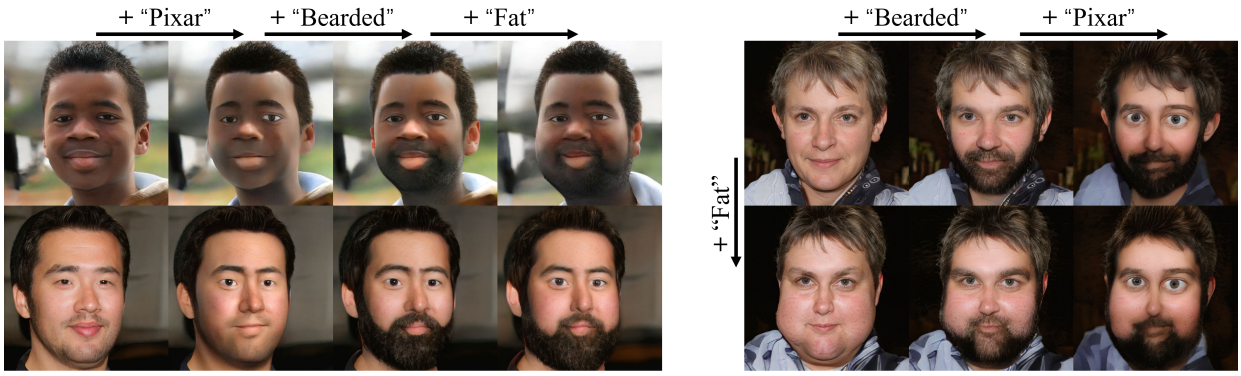
Hence, we can conclude that the shape is mainly controlled by the coarse layers, while the medium and fine layers contribute little to the shape deformation. Similarly, the attribute and the style are dominantly influenced by the medium layers and the fine layers, respectively.

Group choices. Based on the study of multi-level manipulation, we further verify the impact of different choices of layer groups, *i.e.*, , coarse, medium, and fine. To this end, we conduct an additional ablation study in Fig. 8, with a combination of text prompts, *i.e.*, , "Fat", "Bearded Face", and "Pixar". The main part of StyleSDF [41] contains nine linear layers. As can be observed, the division of C: 3, M: 3, and F: 3 can balance all three aspects, *i.e.*, , shape, attribute, and style, while other divisions tend to lack capacity in one respect. For example, the division of C: 2, M: 4, and F: 3 cannot adapt to large shape deformation, the division of C: 4, M: 1, and F: 4 fails to edit an attribute, and the division of C: 3, M: 4, and F: 2 shows little change of style. Hence, we set the coarse/medium/fine layers to 3/3/3 respectively, so as to balance the manipulation of shape, attributes, and style.

Identity loss. ID loss is to preserve the facial identity for intrinsic attributes manipulation. As shown in Fig. 10, with the prompt "blue eyes", the model with ID loss can better keep the facial feature as unchanged as possible while editing the color of the eyes to blue.

E. Extensive Application

Controllable Degree of Manipulation. As described in Sec. III-C, we leverage the hyper-network to predict the parameter offsets for shape deformation, attribution manipulation, and style transfer. Furthermore, the degree of attribute and shape can be linearly controlled by the interpolations done in parameter space. Fig. 9a visualizes the controllability aspect of the shape deformation process based on a coefficient α multiplied by the parameter offsets. Taking "Fat" as an example, with the coefficient rising from 0 to 1, the generated face gradually turns chubby, while a decreasing coefficient can also slim the face. A similar trend can be also found with other shape prompts such as "Long face".



(a) Gradual overlying manipulation (b) cross-overlying manipulation

Fig. 11: **Style mixing**. We showcase visual results of mixing styles of "Fat", "Bearded" and "Pixar", showing the disentanglement of our overlying manipulation.

Fig. 12: **Qualitative results** of cats.

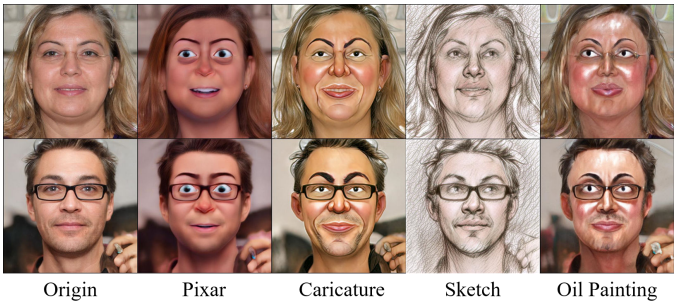


Fig. 13: **Samples of extension** on other 3D-aware GANs.

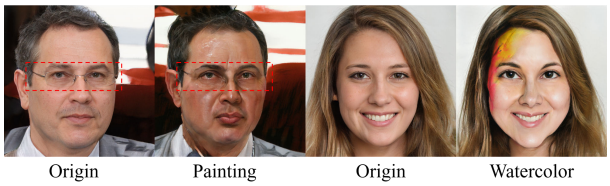


Fig. 14: Failure cases.

Fig. 9b visualizes results of attribute editing controlled by the coefficient α. With the target text prompt of "Old", "Female", and "Silver hair", it is shown that the woman turns old, the male becomes female, and the hair color becomes silver, respectively, as the coefficient α increases.

Style Mixing. Fig. 11 demonstrates the results of style mixing based on our methods. The natural variation of images during mixing of style, attribute and shape shows the good performance on the overlying manipulation. It can handle the editing of different levels without manipulation conflicts while preserving the identity consistency.

Generalization on AFHQ dataset. Although we focus on the task of portrait stylization, we also provide visual results on AFHQ dataset [4] to show the generalization of our methods. We also support the shape deformation, attribute editing and style transfer for cats, as shown in Fig. 12.

Generalization on 3D-aware GANs. Our hypernetwork can also be extend to other 3D-aware models, like EG3D [2]. The stylized results is shown in Fig. 13.

V. Limitations and Future Work

The main limitations of our model is that the hyper-network cannot be expected to well manipulate images according to the text out of the training domain. Besides, it fails in some prompts, *e.g.* , "watercolor", and it is also hard to handle the cases with glasses as shown in Fig. 14.

Meanwhile, we concern about the potential negative influence of our method on society. Our work may be potentially used in the face forgery like DeepFake which may become a threat to personal privacy and reputation. We don't allow any application of our work to malicious behaviors.

VI. Conclusion

In this work, we propose HyperStyle3D, an efficient text-driven method based on 3D-aware GANs for 3D portrait stylization. Specifically, we introduce a hyper-network to predict the offset of the generator parameters with the guidance of the CLIP model. As shown in the experiments, our model achieves high-quality results with regard to style transfer, attribute editing, and shape deformation while avoiding the reliance on rare 3D data. Moreover, we find the style, attribute, and shape are controlled by three separated layer groups in the hyper-network, based on which multi-level overlying manipulations can be realized.

VII. References Section

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, pages 1–21, 2021. 2

[2] 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, 2022. 2, 3, 6, 8, 10

[3] 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

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 10

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 6, 7, 8

[6] 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

[7] 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

[8] 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. 1, 2, 3, 5, 6

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

[10] 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

[11] Fangzhou Han, Shuquan Ye, Mingming He, Menglei Chai, and Jing Liao. Exemplar-based 3d portrait stylization. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 2

[12] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 2

[13] Jialu Huang, Jing Liao, and Sam Kwong. Unsupervised image-to-image translation via pre-trained stylegan2 network. *IEEE Transactions on Multimedia*, pages 1435–1448, 2021. 2

[14] Jialu Huang, Jing Liao, and Sam Kwong. Unsupervised image-to-image translation via pre-trained stylegan2 network. *IEEE Transactions on Multimedia*, 24:1435–1448, 2022. 2

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

[16] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, pages 867–876, 2022. 3

[17] Nikolay Jetchev. Clipmatrix: Text-controlled creation of 3d textured meshes. *arXiv preprint arXiv:2109.12922*, 2021. 3

[18] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Deceive d: Adaptive pseudo augmentation for gan training with limited data. In *NIPS*, pages 21655–21667, 2021. 2

[19] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *ICCV*, pages 13799–13808, 2021. 2

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 1, 2

[21] 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. 2, 3

[22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NIPS*, pages 852–863, 2021. 2, 5

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 5, 6

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 2, 5

[25] 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

[26] 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

[27] 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

[28] 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. 2

[29] 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*, 24:4077–4091, 2022. 2

[30] Shaojie Li, Mingbao Lin, Yan Wang, Fei Chao, Ling Shao, and Rongrong Ji. Learning efficient gans for image translation via differentiable masks and co-attention distillation. *IEEE Transactions on Multimedia*, 25:3180–3189, 2023. 2

[31] 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

[32] 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

[33] Shiguang Liu and Ting Zhu. Structure-guided arbitrary style transfer for artistic image and video. *IEEE Transactions on Multimedia*, 24:1299–1312, 2022. 2

[34] Qi Mao and Siwei Ma. Enhancing style-guided image-to-image translation via self-supervised metric learning. *IEEE Transactions on Multimedia*, 25:8511–8526, 2023. 2

[35] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. *arXiv preprint arXiv:2008.02268*, 2020. 3

[36] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *CVPR*, pages 13492–13502, 2022. 3

[37] 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, 2020. 3

[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] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *CVPR*, pages 10743–10752, 2021. 2

[41] 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. 2, 3, 5, 7, 8, 9

[42] 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

[43] 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. 1, 2, 3, 5

[44] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*, 2020. 2

[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, 3

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

2021. 3
[47] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021. 2
[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. 6
[49] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *CVPR*, pages 18603–18613, 2022. 3
[50] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NIPS*, 2020. 3
[51] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, pages 9243–9252, 2020. 2
[52] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2
[53] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, pages 453–468, 2021. 2
[54] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022. 3
[55] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *arXiv preprint arXiv:2205.15517*, 2022. 2
[56] 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
[57] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *ICML*, pages 9786–9796, 2020. 2
[58] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *CVPR*, pages 3835–3844, 2022. 3
[59] Quan Wang, Sheng Li, Zichi Wang, Xinpeng Zhang, and Guorui Feng. Multi-source style transfer via style disentanglement network. *IEEE Transactions on Multimedia*, pages 1–12, 2023. 2
[60] 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
[61] Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. Deep 3d portrait from a single image. In *CVPR*, pages 7710–7720, 2020. 3
[62] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. In *NIPS*, 2021. 3
[63] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *CVPR*, 2022. 2, 3
[64] Kangxue Yin, Jun Gao, Maria Shugrina, Sameh Khamis, and Sanja Fidler. 3dstylenet: Creating 3d shapes with geometric and texture style variations. In *ICCV*, 2021. 3
[65] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129:3051–3068, 2021. 6
[66] Jichao Zhang, Aliaksandr Siarohin, Yahui Liu, Hao Tang, Nicu Sebe, and Wei Wang. Training and tuning generative neural radiance fields for attribute-conditional 3d-aware face generation. *arXiv preprint arXiv:2208.12550*, 2022. 2, 8
[67] Zhentan Zheng, Jianyi Liu, and Nanning Zheng. P^2-gan: Efficient stroke style transfer using single style image. *IEEE Transactions on Multimedia*, 25:6000–6012, 2023. 2
[68] 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. 2, 3
[69] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d++: End-to-end real-time high-resolution 3d-aware gans for gan inversion and stylization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11502–11520, 2023. 3
[70] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 1, 2
[71] Peiye Zhuang, Oluwasanmi Koyejo, and Alexander G Schwing. Enjoy your editing: Controllable gans for image editing via latent space navigation. *arXiv preprint arXiv:2102.01187*, 2021. 2