Toonify3D: StyleGAN-based 3D Stylized Face Generator

Wonjong Jang
POSTECH
South Korea
wonjong@postech.ac.kr

Yucheol Jung
POSTECH
South Korea
ycjung@postech.ac.kr

Hyomin Kim
POSTECH
South Korea
min00001@postech.ac.kr

Gwangjin Ju
POSTECH
South Korea
gwangjin@postech.ac.kr

Chaewon Son
POSTECH
South Korea
cwson@postech.ac.kr

Jooeun Son
POSTECH
South Korea
jeson@postech.ac.kr

Seungyong Lee
POSTECH
South Korea
leesy@postech.ac.kr

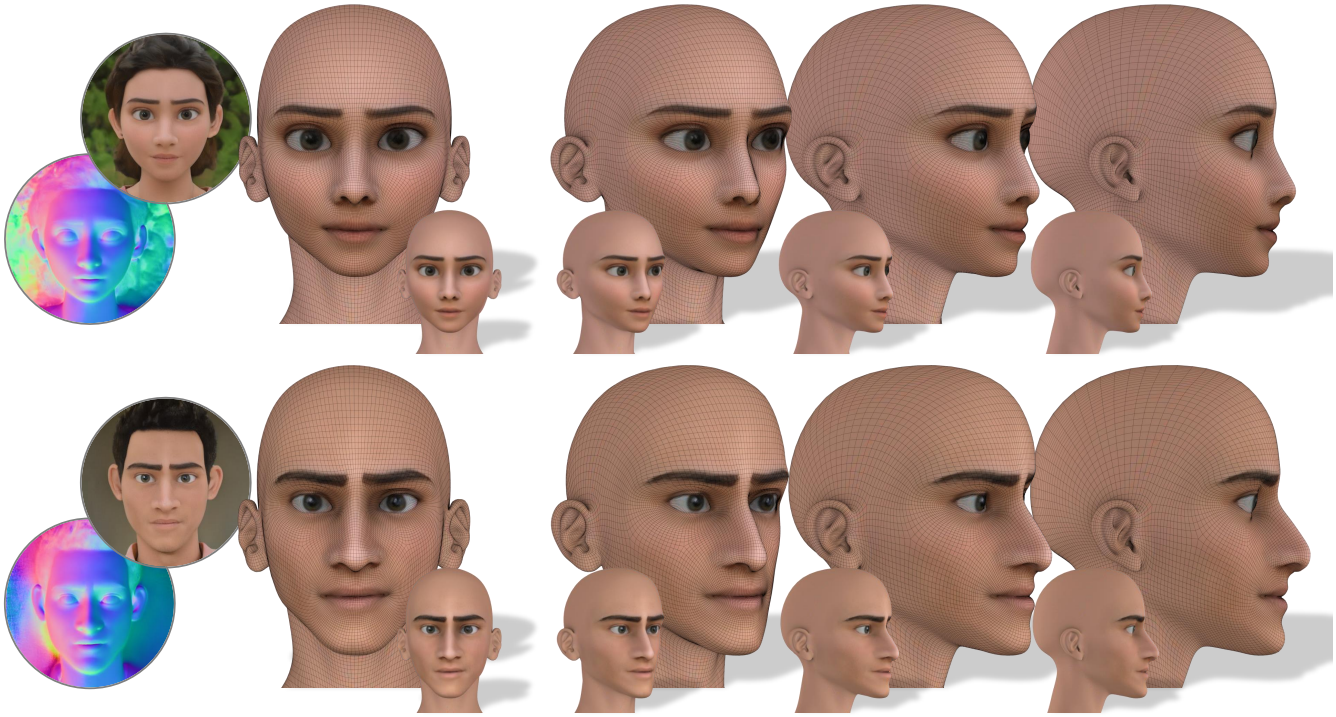

Figure 1: Results of our Toonify3D framework. We turn Toonify [Pinkney and Adler 2020] into a 3D stylized face mesh generator. Given latent codes, our Toonify3D framework generates 3D stylized faces with a shared mesh topology.

ABSTRACT

Recent advances in generative models enable high-quality facial image stylization. Toonify is a popular StyleGAN-based framework that has been widely used for facial image stylization. Our goal is to create expressive 3D faces by turning Toonify into a 3D stylized face generator. Toonify is fine-tuned with a few gradient descent steps from StyleGAN trained for standard faces, and its features would carry semantic and visual information aligned with the features of the original StyleGAN model. Based on this observation, we design a versatile 3D-lifting method for StyleGAN, *StyleNormal*, that regresses a surface normal map of a StyleGAN-generated face using StyleGAN features. Due to the feature alignment between Toonify and StyleGAN, although StyleNormal is trained for regular faces, it can be applied for various stylized faces without additional fine-tuning. To learn local geometry of faces under various illuminations, we introduce a novel regularization term, the *normal consistency loss*, based on lighting manipulation in the GAN latent space. Finally, we present *Toonify3D*, a fully automated framework based on StyleNormal, that can generate full-head 3D stylized avatars and support GAN-based 3D facial expression editing.

SIGGRAPH Conference Papers '24, July 27–August 01, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657480

CCS CONCEPTS

• **Computing methodologies** → *Mesh models.*

KEYWORDS

Toonify, StyleGAN, 3D face stylization

ACM Reference Format:

Wonjong Jang, Yucheol Jung, Hyomin Kim, Gwangjin Ju, Chaewon Son, Jooeun Son, and Seungyong Lee. 2024. Toonify3D: StyleGAN-based 3D Stylized Face Generator. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27–August 01, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657480

1 INTRODUCTION

Recent advances in GANs enable high-quality facial image stylization. Toonify [Pinkney and Adler 2020] is one of the popular approaches for facial stylization based on StyleGAN [Karras et al. 2019]. Toonify generates stylized faces with large and plausible shape exaggerations, but such interesting shapes are only represented as 2D color images. Our goal is to turn Toonify into a 3D stylized face generator by synthesizing 3D shapes that faithfully depict the characteristics of Toonify results.

Synthesizing 3D shapes from stylized facial images is challenging; the reconstruction is ill-posed and data-driven reconstruction is non-trivial due to the difficulty in obtaining ground-truth 3D stylized faces. Previous work often relies on skilled 3D artists to perceptually imitate the shapes of stylized faces. For example, Qiu *et al.* [2021] built a 3D caricature dataset by asking 3D artists to sculpt mesh models that correspond to 2D caricatures. However, manual construction of such dataset is not scalable.

Recent studies on 3D-lifting GANs and 3D-aware GANs provide a fully self-supervised approach for obtaining 3D shapes from GAN-generated images. However, the main objectives for these self-supervisions are 3D-consistent image synthesis, not exactly 3D reconstruction. Then, the self-supervision applied to regular faces [Chan et al. 2022, 2021; Gu et al. 2021; Pan et al. 2020, 2021; Shi et al. 2021] and stylized faces [Abdal et al. 2023; Jin et al. 2022; Wang et al. 2022] may produce smoothed-out or noisy 3D shapes around facial components (see Sec. 5.1 for visual comparison).

In this work, we propose a novel 3D-lifting framework for Toonify to produce 3D stylized faces without ground-truth 3D models or 3D-aware self-supervision. Our key idea is to construct 3D shapes for stylized faces by leveraging prior knowledge on regular face geometries. We define cross-domain StyleGAN features shared by regular and stylized faces, and learn a predictor that maps the cross-domain features to known local 3D geometries of regular faces. The predictor then naturally extends to stylized faces, enabling local 3D geometries to be borrowed from regular faces for stylized facial components.

Previous work [Pakhomov et al. 2021] shows that StyleGAN features effectively convey semantic information in a single domain. Taking a step further, we observe that such characteristics of StyleGAN features can be generalized to stylized domain like Toonify. Toonify is fine-tuned using only a few hundred gradient descent steps starting from a StyleGAN for regular faces. Such weak fine-tuning introduces shape deformation, but not so much deviation in local StyleGAN features (Sec. 3).

Based on this observation, we propose a novel 3D-lifting method, dubbed *StyleNormal*, which learns a mapping from pixel-wise StyleGAN feature vectors to pixel-wise 3D surface normal vectors using a multi-layer perceptron. Thanks to the cross-domain compatibility of StyleGAN features, once trained on regular faces, our StyleNormal can estimate surface normal maps for stylized faces without requiring additional steps to bridge the domain gap. In the training of StyleNormal, we utilize 3D scanned face models to render photo-realistic images and their corresponding surface normal maps.

In the spirit of photometric stereo, we develop a novel regularization term, *normal consistency loss*, which analyzes multiple facial images under various lighting conditions. Using StyleGAN image editing, this term maintains consistent underlying geometry under illumination variation and improves surface normal estimation.

Building upon StyleNormal, we present *Toonify3D*, a fully automated framework that converts Toonify images into 3D full-head meshes with a shared topology. Given a surface normal map from StyleNormal, partial 3D surfaces on the facial region are reconstructed by normal integration. We then perform non-rigid registration of a full-head template mesh to the partial 3D surfaces. The registration enables the resulting 3D shapes from our framework to share a common mesh topology. Using the full-head registration results, we demonstrate automatic construction of 3D facial expressions for 3D stylized avatars using StyleGAN latent space editing.

In summary, our technical contributions are as follows:

- We show StyleGAN features can serve as cross-domain feature descriptor between regular and stylized faces that are useful for predicting local facial geometry.
- We propose *StyleNormal* that can estimate surface normals for stylized faces by leveraging the surface normals of regular faces with the cross-domain StyleGAN feature descriptor.
- We present how to build a synthetic dataset for StyleNormal training and design a normal consistency loss that improves surface normal estimation under various lighting conditions.
- Our Toonify3D framework can generate numerous 3D stylized full-head mesh avatars based on Toonify results, supporting GAN-based 3D facial expression editing.

2 RELATED WORK

Single-view 3D face reconstruction. Estimating a 3D facial shape from a single-view image is a highly ill-posed problem. A traditional approach to solve the ill-posedness is to construct a parametric 3D face model with known parameter distribution [Blanz and Vetter 1999; Li et al. 2017; Paysan et al. 2009]. Parametric 3D face models with *a prior* distributions have been used in various 3D face reconstruction frameworks [Deng et al. 2019; Garrido et al. 2016; Tewari et al. 2017].

However, such models for regular faces do not generalize well to stylized faces, such as cartoon or caricatures. Previous approaches construct separate parametric models for stylized faces by building 3D mesh examples for exaggerated faces. Qui *et al.* [2021] build a 3D caricature dataset by manually sculpting approximately 2,000 3D meshes to mimic the shapes of 2D caricatures. Jung *et al.* [2022] build a neural parametric model for 3D caricatures using the 3D

caricature dataset. Still, constructing a dataset of stylized 3D shapes is laborious and not scalable.

Han *et al.* [2017] build a PCA model for a cartoon by exaggerating regular 3D faces via scaling of surface gradients. However, rule-based exaggeration of regular faces does not guarantee the results represent the target styles well. In our work, we build a 3D face tailored for the target style by building a surface via normal map estimation based on StyleGAN features.

StyleGAN and its features. StyleGAN [Karras et al. 2019, 2020] has been widely used for generating and manipulating photo-realistic facial images. Recent works also demonstrate that the model can be effectively used in the cartoon domain via transfer learning [Gal et al. 2022; Jang et al. 2021; Ojha et al. 2021; Pinkney and Adler 2020; Wang et al. 2022; Yang et al. 2022]. Especially, Toonify [Pinkney and Adler 2020] trains a cartoon image generator by fine-tuning a photo generator with cartoon images. Toonify achieves the blending of photo textures and cartoon structures by mixing StyleGAN features of the two generators.

Recent works [Kim et al. 2022; Nitzan et al. 2022; Pakhomov et al. 2021; Xu et al. 2021; Zhang et al. 2021] leverage StyleGAN features for vision tasks in a single domain, e.g., facial photos. Notably, Segmentation in Style [Pakhomov et al. 2021] and DatasetGAN [Zhang et al. 2021] highlight generalization capabilities of StyleGAN features for the semantic segmentation task in a single domain. Segmentation in Style performs zero-shot semantic segmentation on StyleGAN-generated images by clustering StyleGAN features. DatasetGAN designs a few-shot training scheme for semantic segmentation on StyleGAN-generated images using StyleGAN features as the input for the segmentation network, requiring only a few human annotations.

Polymorphic-GAN [Kim et al. 2022] shows zero-shot segmentation transfer across multiple domains, e.g., facial photos and drawings. Its key idea is to train image generators for other domains using a special architecture that preserves the feature distribution of the original photo model, where the generators are trained by spatially warping StyleGAN feature maps. In our work, we show that establishing semantic correspondence between parent (original) and child (fine-tuned) StyleGAN models does not require training a specialized GAN architecture (Sec. 3). Recently, StyleAlign [Wu et al. 2022] demonstrated cross-domain latent space alignment between parent and child StyleGAN models. Extending this approach, we demonstrate cross-domain feature space alignment of parent and child StyleGAN layers.

Based on cross-domain feature correspondences, we propose StyleNormal that regresses surface normal maps for both regular faces from StyleGAN and stylized faces from Toonify using pixel-wise StyleGAN feature vectors. To our knowledge, our work is the first to utilize cross-domain StyleGAN features for 3D-lifting.

3D geometry acquisition from GANs. 3D shapes can be retrieved from GANs by constructing a 3D-aware image formation model and training the model using the GAN. GAN2Shape [Pan et al. 2020] and LiftedGAN [Shi et al. 2021] learn to generate 3D surfaces from regular GANs using self-supervision that exploits the GAN latent space. Wang *et al.* [2022] build a 3D cartoon generator based on GAN2Shape.

3D-aware face generators based on neural volume rendering [Chan et al. 2022, 2021; Gu et al. 2021; Pan et al. 2021] enable self-supervised training of volumetric 3D face generator from 2D training images. Once trained, the models can generate a volumetric 3D representation for 3D-consistent image synthesis, and a 3D surface can be extracted from the volumetric 3D representation. Recent works on 3D-aware GANs showcase 3D-aware synthesis of stylized faces [Abdal et al. 2023; Jin et al. 2022].

These self-supervised methods for learning stylized 3D generators are advantageous in the sense that they do not require ground-truth geometry. However, relying only on self-supervision may require large-scale stylized face dataset with *exact* camera extrinsic and intrinsic parameters to produce accurate facial geometries. Instead, we train our StyleNormal using a small number (≤ 10) of examples of ground-truth geometry defined on regular face domain. Borrowing the local geometry of high-quality ground-truth 3D regular faces, our framework constructs plausible surfaces for stylized faces.

3 CROSS-DOMAIN STYLEGAN FEATURES

3.1 Cross-domain features of regular and stylized faces

When people observe a stylized facial portrait, they would easily deduce the underlying geometric structure of the face with hardly any effort; while overall proportions of the facial components could be heavily changed, people would be able to associate local regions in the stylized portrait with the real-world counterparts. For example, even if the eyes in the stylized portrait have significantly large size compared to real-world eyes, people can recognize the position, size, and visual characteristics of the eyelids, establishing perceptual correspondence between stylized and realistic eyes.

This observation implies 3D surface for a stylized facial portrait can be constructed by leveraging prior knowledge on regular face geometries. In other words, the same prior on the mapping from local visual appearance to 3D local geometry could be shared among regular and stylized faces, although the arrangements of local geometries can change depending on overall proportions of facial components.

Our key insight is to exploit this observation for constructing 3D surfaces of regular and stylized faces by defining a shared mapping from StyleGAN features to 3D local geometry. In image synthesis, StyleGAN features determine appearances and arrangements of facial components. Furthermore, each StyleGAN layer engages in mutually independent synthesis (e.g., *style-mix* in [Karras et al. 2019]), and their role is maintained even after being fine-tuned for another domain, as shown in *layer-swapping* in [Pinkney and Adler 2020]. Thus, we can hypothesize that features from each StyleGAN layer have common characteristics regardless of the domain.

Based on this hypothesis, we define pixel-wise StyleGAN feature vectors from the StyleGAN feature map pyramid $\mathbf{F}_{pyr} = \{F_1, F_2, ..., F_{18}\}$, where each layer corresponds to the index for the first dimension of the $W+$ space latent code $\mathbf{w} \in \mathbb{R}^{18\times512}$. To obtain a pixel-wise feature vector at a given pixel position, we resize each feature map in the pyramid to the output image size using bi-linear interpolation. We then concatenate and normalize features at the pixel position across all layers. Please refer to the supplementary

document for details on the normalization. In the next section, we show the pixel-wise StyleGAN feature vector is associated with visual appearance, semantic, and 3D geometry of each local region in a consistent way among regular and stylized faces.

3.2 Validation on cross-domain StyleGAN features

To verify the cross-domain compatibility of pixel-wise StyleGAN feature vectors, we perform a nearest neighbor search: for each pixel in the target stylized face image, we associate the pixel with the closest feature distance in the regular face image. Using the nearest neighbor field, we warp the regular face image and its semantic and normal maps, and show the warping results provide plausible constructions of visual, semantic, and geometric information of the stylized face image.

Given feature vectors $\mathbf{v}_r$ and $\mathbf{v}_s$ of regular and stylized faces, respectively, we warp the color image $\mathbf{c}_r$ of the regular face using the nearest neighbor search;

$$\mathbf{c}_s(\mathbf{u}) = \mathbf{c}_r\left(\underset{\mathbf{u}'}{\operatorname{argmin}} \|\mathbf{v}_r(\mathbf{u}') - \mathbf{v}_s(\mathbf{u})\|_2^2\right), \tag{1}$$

where $\mathbf{u}$ and $\mathbf{u}'$ are pixel positions in the stylized and regular face images, respectively. Then, the resulting image $\mathbf{c}_s$ (Fig. 2f bottom) looks very similar to the original color image of the stylized face (Fig. 2b). Warping for segmentation map $\mathbf{s}_r$ and normal map $\mathbf{n}_r$ of the regular face can be performed in a similar way. There are no original segmentation and normal maps to compare in this case, but Fig. 2 shows plausible segmentation and normal maps can be obtained for the stylized face. Fig. 2 also shows warping results using other possible pixel-wise features, such as RGB patch and VGG features [Simonyan and Zisserman 2015], and verifies that our StyleGAN feature vector outperforms them in terms of cross-domain compatibility between regular and stylized faces.

This experimental result has two major implications. First, the StyleGAN features from two different StyleGAN models are compatible if the model is moderately fine-tuned as in Toonify. Second, the StyleGAN features encode local geometric information based on their rich semantic understanding. For instance, the segmentation map from StyleGAN features shows that all stylized facial geometries come from their respective facial components in the regular face. In addition, notably, the normal map obtained by warping for the stylized face shows convincing geometry for the exaggerated facial components in Fig. 2.

4 NORMAL-BASED 3D LIFTING FOR STYLEGAN

In this section, we propose a novel 3D-lifting method for StyleGAN, which we call *StyleNormal*, that achieves versatile 3D-lifting for both regular faces from StyleGAN and stylized faces from Toonify. Our StyleNormal is a small neural network designed as a 3D-lifting add-on for StyleGAN2 [Karras et al. 2020] and learns the mapping from a pixel-wise StyleGAN feature vector defined in Sec. 3 to a surface normal vector. Contrary to other alternatives such as depth, which represent a shape in 3D global coordinates, surface normal only specifies 3D local geometry that can be directly associated with

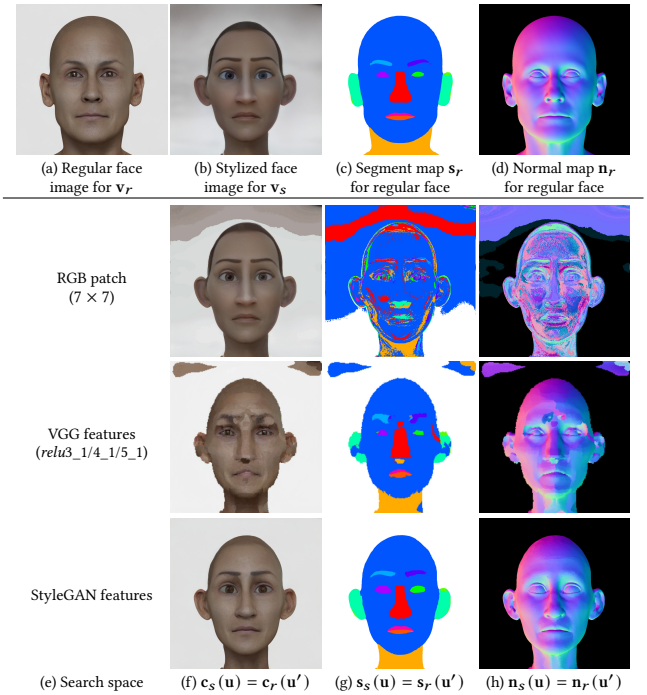

Figure 2: Results of nearest-neighbor search in the StyleGAN feature space. We re-arrange colors, segmentation labels, and surface normals from a regular face by performing nearest-neighbor search with pixel-wise StyleGAN feature vector. Due to the cross-domain compatibility of the feature vector, simple greedy assignment based on feature distance produces plausible warping results for visual appearance, segmentation map, and 3D normal map of the stylized face. Inputs: ©3DScanStore

local visual appearance encoded in StyleGAN features, as demonstrated in Fig. 2. We train StyleNormal using regular faces only, but thanks to the cross-domain compatibility of StyleGAN features discussed in Sec. 3, the trained StyleNormal can be readily applied to stylized faces. Leveraging the 3D-lifting result of StyleNormal, our *Toonify3D* framework can generate 3D full-head meshes with a shared mesh topology useful for real-world applications.

4.1 Synthetic dataset for StyleNormal

To train StyleNormal, which transforms a pixel-wise StyleGAN feature vector $\mathbf{v} \in \mathbb{R}^{6080}$ to its corresponding surface normal vector $\mathbf{n} \in \mathbb{R}^3$, we require a dataset comprising of paired $(\mathbf{v}, \mathbf{n})$ samples. We build this dataset using 3D scanned human data from 3DScanStore [3DScanStore 2023] under natural environmental light conditions. We render color and normal images of ten 3D scanned humans using Blender [Blender 2023], under six distinct environmental lighting conditions, resulting in total 60 colored images and 10 surface normal maps. During training, we ignore pixels where surface normal maps have not been rendered.

Since the input of StyleNormal is a pixel-wise StyleGAN feature vector $\mathbf{v}$, we extract StyleGAN features from the color images in the dataset using GAN inversion [Richardson et al. 2021; Roich et al.

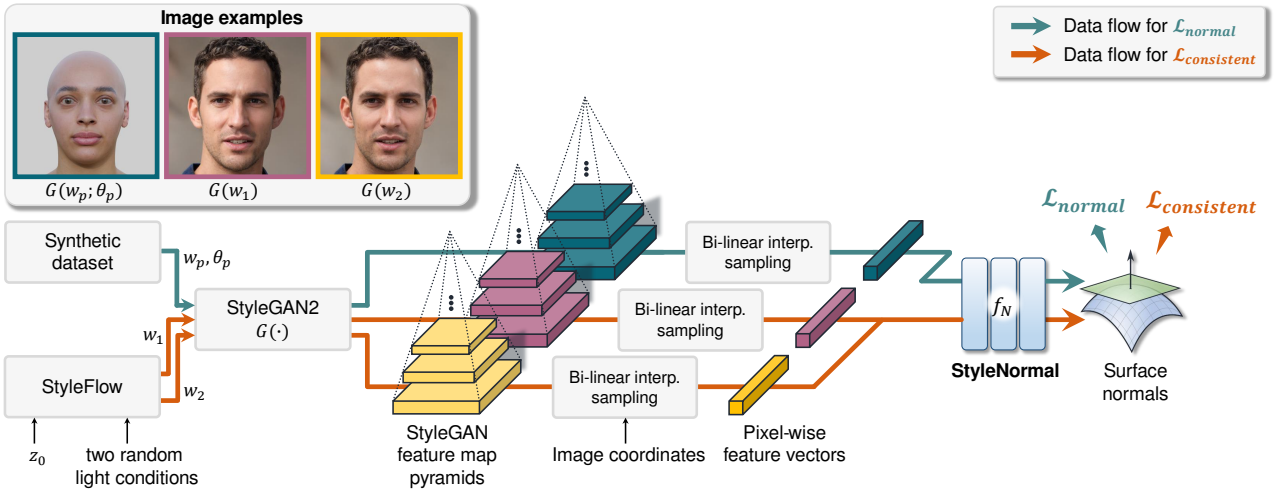

Figure 3: Overview of StyleNormal training. At each iteration, we sample one pair of latent code w_p and network parameters θ_p obtained from GAN inversion on a synthetic image in our dataset. A pixel-wise feature vector extracted from the pair is fed to StyleNormal to calculate $\mathcal{L}_{normal}$. For regularization, we sample two additional pixel-wise feature vectors from StyleGAN by sampling a random latent code z_0 from Gaussian distribution and applying latent-space editing for two random lighting conditions. The two feature vectors are used to calculate $\mathcal{L}_{consistent}$.

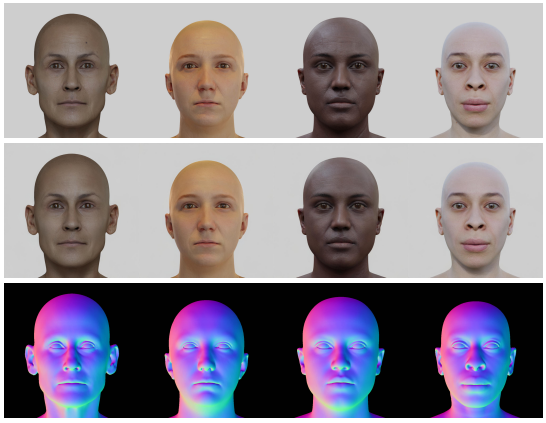

Figure 4: Examples of our synthetic dataset constructed with 3DScanStore data [3DScanStore 2023]. Rendered RGB images (top row), GAN projection via PTI [Roich et al. 2022] (middle), and rendered surface normals (bottom). Note that the facial geometries are almost unchanged after GAN inversion. Inputs: ©3DScanStore

2022; Tov et al. 2021; Zhu et al. 2020]. In this process, we employ pivotal tuning GAN inversion [Roich et al. 2022], a technique that achieves accurate reconstruction of the input image while enforcing editable GAN latent space (Fig. 4). After performing the GAN inversion, we construct pixel-wise StyleGAN feature vectors as decribed in Sec. 3. As the pivotal tuning faithfully maintains the input appearance and in-domain properties, we can align pixel-wise StyleGAN features with the rendered surface normal maps.

4.2 Training StyleNormal

Our StyleNormal is a neural network that infers a surface normal map for the facial region in a StyleGAN-generated image. It is constructed as a four-layer multi-layer perceptron (MLP) regressor f_N. The overall training process is illustrated in Fig. 3.

In training phase, we randomly sample image coordinates $\mathbf{u}$ inside the head region of the rendered surface normal map. StyleNormal f_N is updated using the following loss function:

$$\mathcal{L}_{normal} = \mathbb{E}_{(\mathbf{v},\mathbf{n})\sim p_{data}}[\|f_N(\mathbf{v}(\mathbf{u})) - \mathbf{n}(\mathbf{u})\|_1], \quad (2)$$

where $\mathbf{v}(\mathbf{u})$ and $\mathbf{n}(\mathbf{u})$ are pixel-wise StyleGAN feature vector and ground-truth surface normal vector sampled at image coordinates $\mathbf{u}$, respectively.

To effectively estimate surface normal maps under diverse lighting conditions, it is advantageous to learn surface normal estimation for a range of lighting scenarios. To facilitate this, we propose a *normal consistency loss*. Inspired by photometric stereo, which analyzes multiple images of an object under varying illumination, we constrain our StyleNormal to estimate a consistent surface normal map for a single face when the light direction changes. To synthesize diverse illuminations, we adopt StyleFlow [Abdal et al. 2021], a recent attribute-conditioned latent space exploration method, that supports illumination editing while preserving the underlying geometry. This regularization ensures that StyleNormal is exposed to multiple illumination conditions for a single unseen geometry, thereby enhancing its generalization capability.

In the preprocessing step, we randomly sample a set of latent codes $\mathbf{z}_0$ from Gaussian distribution. For each sampled latent code, we obtain five illumination-edited latent codes $\mathbf{w}$ by applying StyleFlow with five distinct illumination direction presets; front, left, right, above, and below. Using these latent codes $\mathbf{w}$, we generate five StyleGAN feature map pyramids ($G(\mathbf{w})$ in Fig. 3) that introduce lighting variations in StyleGAN feature vectors used for training.

During each training iteration, we randomly select two light-adjusted StyleGAN feature map pyramids $G(w_1)$ and $G(w_2)$, and then sample pixel-wise feature vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ from $G(w_1)$ and $G(w_2)$ at the same image coordinates $\mathbf{u}'$, respectively. Our *normal*

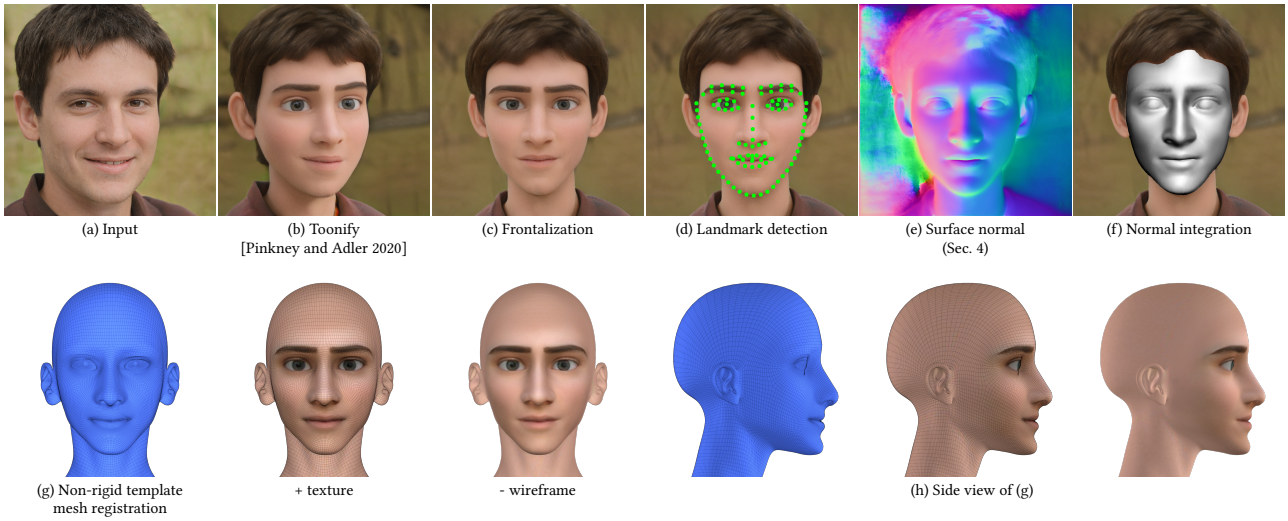

Figure 5: Overall process for our full-head 3D stylized face generation.

Figure 6: Results of StyleNormal from Toonify outputs of various styles. Without requiring additional training, StyleNormal generates plausible 3D normals for multiple domains.

consistency loss is then formulated as

$$\mathcal{L}_{consistent} = \mathbb{E}[\|f_N(\mathbf{v}_1(\mathbf{u}')) - f_N(\mathbf{v}_2(\mathbf{u}'))\|_1], \tag{3}$$

which enforces the estimated normals $f_N(\mathbf{v}_1)$ and $f_N(\mathbf{v}_2)$ to be the same, as they should represent the same underlying local geometry independently from adjusted lighting.

To summarize, the full objective function for training StyleNormal f_N is as follows:

$$\mathcal{L}_{f_N} = \mathcal{L}_{normal} + \lambda_{reg}\mathcal{L}_{consistent}, \tag{4}$$

where λ_{reg} is a hyper-parameter for normal consistency loss.

Once StyleNormal has been trained, it can be applied to a variety of Toonify models, each trained with different stylized face datasets, without requiring any additional training (Fig. 6).

4.3 Full-head 3D Stylized Face Generation

By applying StyleNormal to a Toonify result, we can obtain a surface normal map that describes the 3D surface geometry of a stylized face. However, surface normal maps alone have limited applications. We therefore design a fully automated framework, *Toonify3D*, to obtain full-head mesh models from Toonify results. Fig. 5 illustrates the overall process for generating a full-head 3D stylized mesh model.

We first transform the pose of the input image into a frontal pose using latent space editing [Shen et al. 2020]. We then extract a surface normal map using our StyleNormal and convert the surface normal map into a depth map by performing normal integration [Hertzmann and Seitz 2005] on the facial region. We also detect facial landmarks using [Jin et al. 2021] and then non-rigidly fit a full-head template mesh to the partial surface with facial landmarks using [Amberg et al. 2007]. Implementation details are discussed in the supplementary document.

5 EXPERIMENTS

Our Toonify3D framework[1] generates a 3D full-head stylized avatar by taking a real-world photo or a Toonify-generated image as the input (Figs. 11 and 15). The resulting avatar can be used in real-world applications by attaching diverse 3D assets depending on user preferences (Fig. 14).

Running time. The entire process of Toonify3D takes about 3 minutes using Intel Xeon Gold 6226R CPU and NVIDIA Quadro RTX 8000 GPU. StyleNormal inference and surface normal integration take 1.2 miliseconds and 15 seconds, respectively. The rest of the time is spent on pre-processing, post-processing, and non-rigid registration.

5.1 Qualitative comparison

We conduct a qualitative evaluation of our results compared to state-of-the-art methods for 3D-aware GANs, from which a 3D

[1] Our code is available at https://github.com/wonjongg/Toonify3D

(a) LiftedGAN [Shi et al. 2021] (b) Ours (c) EG3D [Chan et al. 2022] (d) Ours

Figure 7: Qualitative comparison to state-of-the-art methods for 3D-lifting GAN [Shi et al. 2021] and 3D-aware GAN [Chan et al. 2022] for regular face domain. The images (a) and (c) are taken from figures of respective works. For side-by-side comparison of geometry quality, we apply GAN-inversion to the regular face images and visualize our results as shading images. Our surface shows cleaner geometry for facial components.

surface can be extracted. For further qualitative comparisons with 3D reconstruction methods and a 3D diffusion model, please refer to the supplementary document.

For regular faces, we compare the visual quality of 3D surfaces extracted by our method with that of LiftedGAN [Shi et al. 2021] and EG3D [Chan et al. 2022] (Fig. 7). LiftedGAN performs unsupervised 3D-lifting for 2D GANs, and EG3D is a GAN for a volumetric radiance field. To facilitate side-by-side comparison, we first perform GAN inversion [Roich et al. 2022] to obtain StyleGAN features of input images and then apply our StyleNormal to these extracted features.

We also visually compare 3D surfaces for stylized faces in Fig. 10, which includes diverse state-of-the-art 3D-aware GANs capable of yielding 3D surfaces. For EG3D [Chan et al. 2022] and Dr.3D [Jin et al. 2022], we fine-tune each model with cartoon images and then use GAN inversion on input images to obtain 3D geometries. Further details on fine-tuning and the GAN inversion are available in the supplementary document. For E3DGE [Lan et al. 2023] and 3DAvatarGAN [Abdal et al. 2023], both of which consider cartoon face domain, we use examples provided in their works for direct comparison.

The 3D surfaces from previous 3D-aware GANs tend to produce rough geometries based on estimated depths or volumetric representations. Although these rough shapes from previous methods are suitable for 3D-consistent image synthesis, their utility in creating 3D mesh models for avatars is limited, particularly due to the lack of fine shape features, such as those around eyes and nose. In contrast, our method produces clear facial surfaces that enable more accurate facial identity recognition based on geometry. As a result, these surfaces are better suited for visually pleasing non-rigid template mesh registration.

5.2 Quantitative comparison

To ensure reliable surface normal map generation for Toonify results, it is crucial to verify that StyleNormal is accurately trained to estimate correct surface normal maps for regular faces. We evaluate the accuracy of StyleNormal for regular faces in Table 1. The errors in estimated surface normals for regular faces are measured using our test set, which comprises 60 rendered facial images and the corresponding surface normal maps obtained using ten 3D scanned faces [3DScanStore 2023] under six different illuminations.

Compared to existing frameworks that require large-scale training data, such as pix2vertex [Sela et al. 2017], SfSNet [Sengupta et al. 2018], and cross-modal network [Abrevaya et al. 2020], our StyleNormal achieves more accurate surface normal estimation with a limited set of training data. StyleNormal leverages StyleGAN features, which contain rich semantic and geometric information, thus effectively addressing the ill-posedness of normal estimation.

5.3 Ablation study

Effect of $\mathcal{L}_{consistent}$. We first evaluate the effect of the loss on the accuracy of the estimated normal maps for regular faces (Table 2). We report estimation errors compared to ground-truth normal maps, using the test set of our 3D scanned human face dataset. Inclusion of $\mathcal{L}_{consistent}$ decreases the error by regularizing StyleNormal with various lighting conditions.

We then evaluate the effect of this loss on the robustness of estimated 3D stylized normal maps against light variations in Toonify results. We create a hundred cartoon images using Toonify and apply five different light manipulations using StyleFlow [Abdal et al. 2021]. We then report the mean angular errors between the five estimated surface normal maps and their average. The $\mathcal{L}_{consistent}$ loss that enforces insensitivity to lighting variations in regular faces consequently makes StyleNormal robust against diverse stylized lighting conditions in Toonify results. Fig. 13 visualizes the effect of $\mathcal{L}_{consistent}$ on StyleNormal results, where 3D facial geometries are more faithfully recovered regardless of stylized lighting effects.

Training data size. In Table 3, we explore the effect of training data size on the accuracy of surface normal estimation. Our results show that increasing the number of identities used for training from 1 to 10 leads to a decrease in error. However, the performance gain is marginal and tends to converge. Since our StyleNormal is trained in a pixel-wise manner and a single image contains about 300,000 pixels in the facial region, we can effectively train StyleNormal with large amount of pixel-wise paired data even with a small number of face images.

Please refer to the supplementary document for additional ablation studies on *number of layers* and *input feature layer selection*.

5.4 Generalization to diverse styles

We verify the generalization power of our StyleNormal to diverse stylized faces. To synthesize images for diverse domains, we adapt StyleGAN-NADA [Gal et al. 2022], which presents text-driven domain adaptation of styleGAN by leveraging the semantic power of large-scale Contrastive-Language-Image-Pre-training (CLIP) [Radford et al. 2021] model. We extract StyleGAN features from StyleGAN-NADA models and then estimate surface normal maps using our StyleNormal. In Fig. 8, we observe that our StyleNormal can successfully estimate surface normal maps from diverse face stylization. Please refer to the supplementary document for more results.

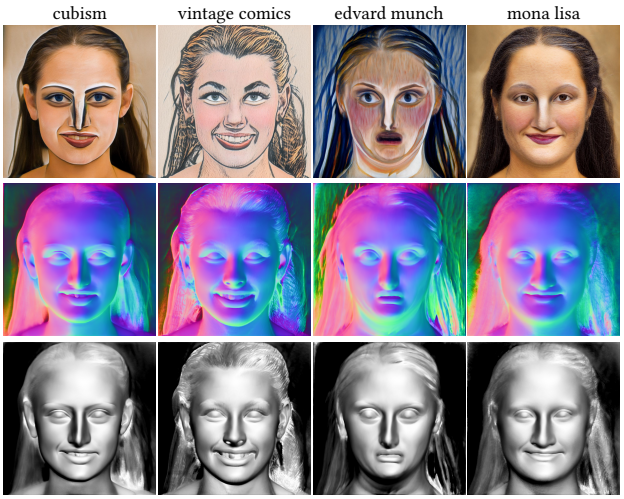

Figure 8: Results of StyleNormal on various styles from StyleGAN-NADA. We annotate the style name above each result. For each style, three images are presented: (top) input image, (middle) StyleNormal result, and (bottom) shading of StyleNormal result for visualization.

5.5 Application to GAN-based 3D shape editing

We introduced an automated pipeline that facilitates transformation of Toonify's 2D outputs into 3D full-head stylized faces. Using this pipeline as the bridge between 2D GANs and 3D meshes, we can also achieve 3D shape editing based on semantic manipulation in GAN latent space. Fig. 12 demonstrates 3D facial expression editing based on 2D expression editing with InterFaceGAN [Shen et al. 2020]. Such 3D modeling based on rich 2D appearance model could open up a wide set of tool-kits for authoring stylized 3D faces.

5.6 Limitations

Toonify bias. Our framework relies on a Toonify-based backbone for face stylization. Toonify is trained using a limited number of cartoon characters, thus suffering with the model bias toward certain character styles. We expect that this limited diversity can be resolved by training Toonify with larger cartoon dataset or applying various text-conditioned styles from StyleGAN-NADA.

Subjects with eyeglasses. Our acquisition of depth maps from estimated normal maps assumes that the facial region is continuous. In case where Toonify results include eyeglasses, the presence of eyeglasses introduces discontinuities in the surface normal maps, particularly around the frames and lenses. StyleNormal results on subjects wearing eyeglasses are shown in Fig. 9.

6 CONCLUSION

In this work, we presented Toonify3D, a novel 3D full-head stylized face generation framework based on Toonify. We experimentally demonstrated the compatibility of Toonify features with original StyleGAN features. Our method, which adapts the local 3D geometry of regular faces to stylized faces based on cross-domain feature compatibility, conceptually aligns with the perception of 3D geometry in stylized faces. Building on this concept, we developed

Figure 9: Limitations. If the input subject wears eyeglasses, strong discontinuity in normal values near the glasses may hinder reliable surface normal integration.

Table 1: Quantitative evaluation of normal estimation for regular faces on our test set comprising 3D scanned data [3DScanStore 2023]. Reported mean angular errors are the average of angles between the estimated and GT 3D normals in degrees. Percentages of pixels within different error thresholds are reported on the right.

Method (Dataset volume)	Mean±std	< 20°	< 25°	< 30°
pix2vertex [2017] (large-scale)	34.51±2.25	54.0%	66.1%	72.2%
SfSNet [2018] (large-scale)	17.32±1.60	70.0%	78.1%	83.8%
cross-modal [2020] (large-scale)	16.66±1.46	73.4%	84.0%	90.1%
Ours (few-shot)	**9.57±1.26**	**91.8%**	**95.1%**	**96.9%**

Table 2: Ablation study on normal consistency loss. The error metric is the same as in Table 1.

Evaluation	w/o $\mathcal{L}_{consistent}$	full
Accuracy for regular face normal	9.85±1.41	**9.57±1.26**
Robustness under light variation	3.18±0.79	**1.85±0.61**

Table 3: Ablation study on the number of identities (N) used for training. As mentioned in Sec. 4.1, we used six rendered images per identity for training. The error metric is the same as in Table 1.

Number of identities	Mean±std	< 10°	< 15°	< 20°
$N = 1$ (6 shots)	10.52±0.88	60.0%	81.3%	90.2%
$N = 3$ (18 shots)	9.63±1.37	66.3%	84.6%	91.8%
$N = 5$ (30 shots)	9.59±1.46	66.7%	84.3%	91.7%
$N = 10$ (60 shots)	9.57±1.26	67.3%	84.7%	91.8%

StyleNormal, a 3D-lifting add-on for both StyleGAN and Toonify. Using StyleNormal as the bridge between a 2D generative model and 3D modeling, creative 3D face authoring toolkits could be built upon rich 2D GAN latent spaces.

ACKNOWLEDGMENTS

This work was supported by the NRF grant (RS-2023-00280400) and IITP grants (AI Graduate School Program, 2019-0-01906; AI Innovation Hub, 2021-0-02068) funded by Korea government (MSIT).

REFERENCES

3DScanStore. 2023. *3DScanStore.* https://www.3dscanstore.com

Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 2023. 3DAvatarGAN: Bridging Domains for Personalized Editable Avatars. In *Proc. CVPR.*

Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. 2021. StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows. *ACM Trans. Graph.* (2021).

Victoria Fernandez Abrevaya, Adnane Boukhayma, Philip H.S. Torr, and Edmond Boyer. 2020. Cross-Modal Deep Face Normals With Deactivable Skip Connections. In *Proc. CVPR.*

Brian Amberg, Sami Romdhani, and Thomas Vetter. 2007. Optimal step nonrigid ICP algorithms for surface registration. In *Proc. CVPR.*

David Beniaguev. 2022. Synthetic Faces High Quality (SFHQ) dataset. https://github.com/SelfishGene/SFHQ-dataset

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proc. SIGGRAPH.*

Blender. 2023. *Blender - a 3D modelling and rendering package.* http://www.blender.org

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proc. CVPR.*

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

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set. In *Proc. CVPR Workshops.*

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph.* (2022).

Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of personalized 3D face rigs from monocular video. *ACM Trans. Graph.* (2016).

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2021. StyleNeRF: A style-based 3D aware generator for high-resolution image synthesis. In *Proc. ICLR.*

Xiaoguang Han, Chang Gao, and Yizhou Yu. 2017. DeepSketch2Face: a deep learning based sketching system for 3D face and caricature modeling. *ACM Trans. Graph.* (2017).

Aaron Hertzmann and Steven M Seitz. 2005. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *IEEE TPAMI* (2005).

Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. 2021. StyleCariGAN: Caricature generation via StyleGAN feature map modulation. *ACM Trans. Graph.* (2021).

Haibo Jin, Shengcai Liao, and Ling Shao. 2021. Pixel-in-Pixel Net: Towards Efficient Facial Landmark Detection in the Wild. *IJCV* (2021).

Wonjoon Jin, Nuri Ryu, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. 2022. Dr.3D: Adapting 3D GANs to Artistic Drawings. In *Proc. ACM SIGGRAPH Asia.*

Yucheol Jung, Wonjong Jang, Soongjin Kim, Jiaolong Yang, Xin Tong, and Seungyong Lee. 2022. Deep deformable 3d caricatures with learned shape control. In *Proc. ACM SIGGRAPH.*

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR.*

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR.*

Seung Wook Kim, Karsten Kreis, Daiqing Li, Antonio Torralba, and Sanja Fidler. 2022. Polymorphic-GAN: Generating aligned samples across multiple domains with learned morph maps. In *Proc. CVPR.*

Yushi Lan, Xuyi Meng, Shuai Yang, Chen Change Loy, and Bo Dai. 2023. E3DGE: Self-Supervised Geometry-Aware Encoder for Style-based 3D GAN Inversion. In *Proc. CVPR.*

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* (2017).

Yotam Nitzan, Rinon Gal, Ofir Brenner, and Daniel Cohen-Or. 2022. Large: Latent-based regression through gan semantics. In *Proc. CVPR.*

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

Daniil Pakhomov, Sanchit Hira, Narayani Wagle, Kemar E Green, and Nassir Navab. 2021. Segmentation in style: Unsupervised semantic image segmentation with stylegan and clip. *arXiv preprint arXiv:2107.12518* (2021).

Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. 2020. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. *arXiv preprint arXiv:2011.00844* (2020).

Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. 2021. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. *NeurIPS* (2021).

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *Proc. AVSS.*

Justin NM Pinkney and Doron Adler. 2020. Resolution dependent GAN interpolation for controllable image synthesis between domains. In *Proc. NeurIPS workshop on Machine Learning for Creativity and Design.*

Yuda Qiu, Xiaojie Xu, Lingteng Qiu, Yan Pan, Yushuang Wu, Weikai Chen, and Xiaoguang Han. 2021. 3DCaricShop: A dataset and a baseline method for single-view 3d caricature face reconstruction. In *Proc. CVPR.*

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

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proc. CVPR.*

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2022. Pivotal Tuning for Latent-based Editing of Real Images. *ACM Trans. Graph.* (2022).

Matan Sela, Elad Richardson, and Ron Kimmel. 2017. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proc. ICCV.*

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. 2018. SfSNet: Learning shape, reflectance and illuminance of faces in the wild. In *Proc. CVPR.*

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE TPAMI* (2020).

Yichun Shi, Divyansh Aggarwal, and Anil K Jain. 2021. Lifting 2d stylegan for 3d-aware face generation. In *Proc. CVPR.*

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR.*

Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proc. ICCV Workshops.*

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an encoder for stylegan image manipulation. *ACM Trans. Graph.* (2021).

Hao Wang, Guosheng Lin, Steven CH Hoi, and Chunyan Miao. 2022. 3D cartoon face generation with controllable expressions from a single GAN image. *arXiv preprint arXiv:2207.14425* (2022).

Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. 2022. StyleAlign: Analysis and Applications of Aligned StyleGAN Models. In *Proc. ICLR.*

Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. 2021. Generative hierarchical features from synthesizing images. In *Proc. CVPR.*

Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2022. Pastiche Master: Exemplar-based high-resolution portrait style transfer. In *Proc. CVPR.*

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. 2021. DatasetGAN: Efficient labeled data factory with minimal human effort. In *Proc. CVPR.*

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-domain gan inversion for real image editing. In *Proc. ECCV.*

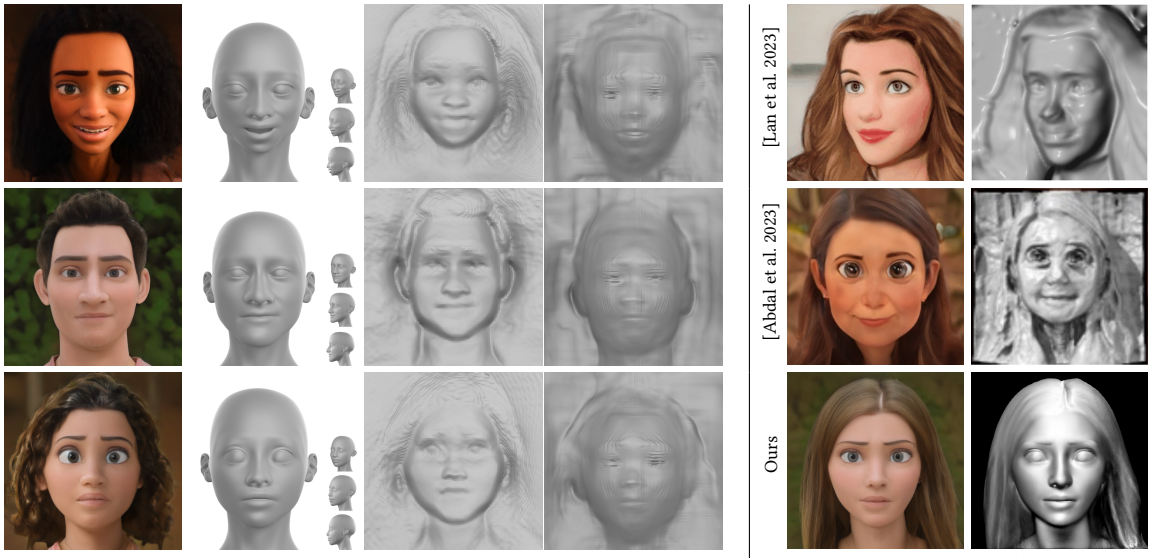

Figure 10: Qualitative comparison to state-of-the-art 3D-aware GANs applied to stylized face domain. We fine-tuned EG3D [Chan et al. 2022] and Dr.3D [Jin et al. 2022] on cartoon face dataset and performed GAN inversion to acquire facial geometries. For E3DGE [Lan et al. 2023] and 3DAvatarGAN [Abdal et al. 2023] that already consider cartoon face domain, examples provided in their works are used for comparison. 3D-aware GANs generally produce flat or blurred shapes while our results create clear shapes for facial components.

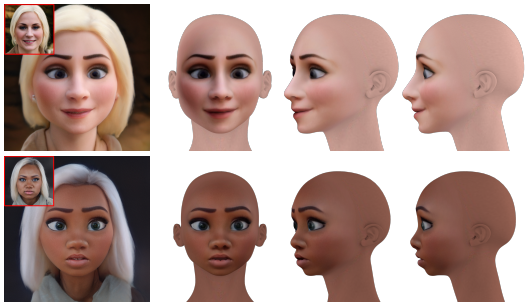

Figure 11: Results on real-world photos. Input photos are from SFHQ dataset [Beniaguev 2022].

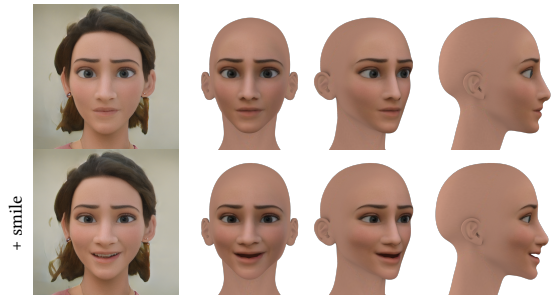

Figure 12: GAN-based 3D expression editing. By applying latent-space semantic editing on a 2D image, we can obtain the corresponding 3D shape with edited expression.

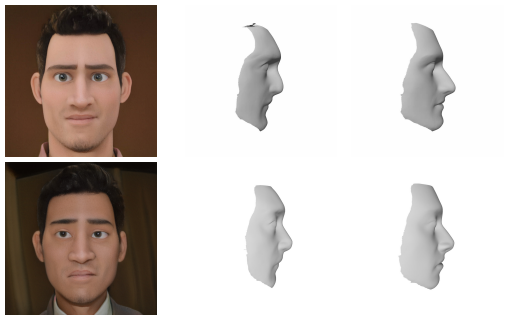

Figure 13: Effect of normal consistency loss on the 3D shapes from Toonify results.

Figure 14: Using our 3D full-head results, users can attach diverse 3D assets, such as hair and clothes, depending on their preferences.

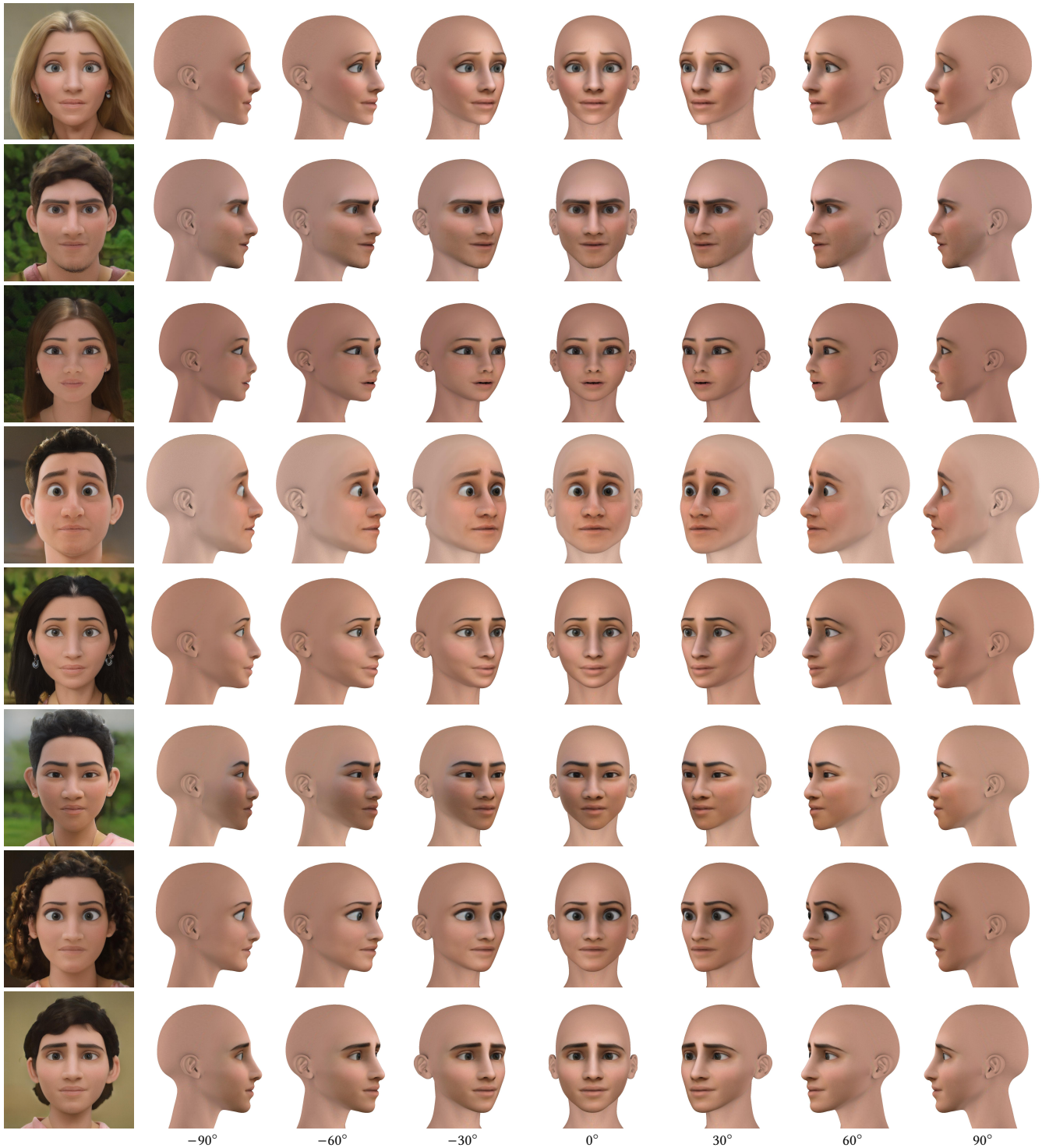

Figure 15: Multi-view visualization of our results. More results can be found in our supplementary document.