Your3dEmoji: Creating Personalized Emojis via One-shot 3D-aware Cartoon Avatar Synthesis

Shiyao Xu*
WICT, Peking University
Beijing, China
xusy@stu.pku.edu.cn

Lingzhi Li
Alibaba Group
Beijing, China
llz273714@alibaba-inc.com

Li Shen
Alibaba Group
Beijing, China
jinyan.sl@alibaba-inc.com

Yifang Men
Alibaba Group
Beijing, China
myf272609@alibaba-inc.com

Zhouhui Lian†
WICT, Peking Univeristy
Beijing, China
lianzhouhui@pku.edu.cn

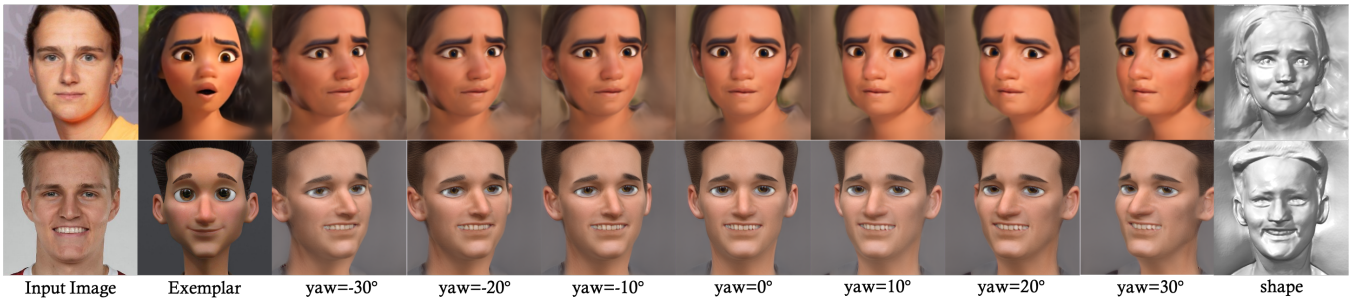

Figure 1: Given a portrait and a style examplar, our method can generate the cartoonized avatar with true-to-life 3D geometry.

ABSTRACT

Creating cartoon-style avatars has drawn growing attention recently, however previous methods only learn face cartoonization in the 2D image level. In this paper, we propose a novel 3D generative model to translate a real-world face image into its corresponding 3D avatar with only a single style example provided. To bridge the gap between 2D real faces and 3D cartoon avatars, we leverage the state-of-the-art StyleGAN and its style-mixing property to produce a 2D paired cartoonized face dataset. We then finetune a pretrained 3D GAN with the pair data in a dual-learning mechanism to get the final synthesized 3D avatar. Furthermore, we analyze the latent space of our model, enabling manual control in what degree a style is applied. Our model is 3D-aware in the sense and also able to do attribute editing, such as smile, age, etc directly in the 3D domain. Experimental results demonstrate that our method can produce high-fidelity cartoonized avatars with true-to-life 3D geometry.

CCS CONCEPTS

• **Computing methodologies → 3D imaging**.

SA '22 Technical Communications, December 6–9, 2022, Daegu, Republic of Korea

ACM ISBN 978-1-4503-9465-9/22/12…$15.00
https://doi.org/10.1145/3550340.3564220

KEYWORDS

Style Transfer, 3D-aware generation

ACM Reference Format:
Shiyao Xu*, Lingzhi Li, Li Shen, Yifang Men, and Zhouhui Lian†. 2022. Your3dEmoji: Creating Personalized Emojis via One-shot 3D-aware Cartoon Avatar Synthesis. In *SIGGRAPH Asia 2022 Technical Communications (SA '22 Technical Communications), December 6–9, 2022, Daegu, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3550340.3564220

1 INTRODUCTION

With the popularity of various short video apps, portrait cartoonization has attracted great interest in the vision and graphics communities. However, current methods can only generate pseudo-3d results without modeling the underlying 3d geometry. This drawback makes them more like the synthesis results obtained by applying some filters to the 2D images, lacking view-consistency in nature. How to produce real 3D avatars with geometry consistency and achieve attribute editing is a meaningful but challenging task.

In the last few years, a number of approaches for exemplar-based portrait stylization have been reported [Men et al. 2022], [Yang et al. 2022], [Chong and Forsyth 2021], achieving impressive results. However, they all operate in the 2D domain and thus produce results lacking 3D geometry and consistency of the specific input exemplar, such as a Pixar cartoon character. On the contrary, we make our model that is fed with a single-view portrait more 3D-aware by leveraging a state-of-the-art 3DGAN model, i.e., [Chan et al. 2022],

* Work done during an internship at Alibaba Group. † Corresponding author.

which is fine-tuned by the paired 2D cartoonized portraits generated by StyleGAN2 [Karras et al. 2020].

Most recent works on 3DGANs [Gu et al. 2021], [Or-El et al. 2022], [Chan et al. 2022] chose to combine StyleGAN [Karras et al. 2019] [Karras et al. 2020] [Karras et al. 2021] with the neural rendering module to synthesize high-quality and high-resolution images with a good multi-view consistency. They can be trained on collections of single-view 2D portraits without any extra 3D geometry information for supervision, achieving a good balance between image quality and computational cost. Since we want to synthesize 3D-aware stylized results, an intuitive idea is to extend the stylization operations from 2DGANs to 3DGANs. However, it is tough to achieve this goal. First, existing 3DGAN models typically need to be trained on tens of millions of images, and their neural rendering modules are quite computationally expensive making them difficult for synthesizing high-resolution images. Second, fine-tuning a pre-trained 3DGAN in the target domain cannot work well without sufficient data, while this paper aims to stylize in-the-wild target images by using only one style exemplar provided. Finally, taking a state-of-the-art 3D generator EG3D [Chan et al. 2022] as an example, it needs the information of camera pose as the condition to decouple pose-correlated attributes during training, which is difficult to extract from the single exemplar cartoon image.

To tackle the above-mentioned challenges, we propose a novel dual architecture (see Fig. 2) that uses the intermediate images stylized by a 2D StyleGAN generator as guidance, to help supervise the 3D-aware generator to produce 3D geometry results. Inspired by the previous one-shot img2img translation works [Chong and Forsyth 2021], [Yang et al. 2022], we destylize the input exemplar image to a real-face style vector and implement style mixing to generate a set of paired data to finetune the StyleGAN model. The wild image is then embedded to both a latent vector of 2D domain w_{2d} and a latent vector 3D domain w_{3d}. Afterwards, the former latent vector is fed into the fine-tuned StyleGAN generator to generate the stylized image, forming a image pair with the original input image. Finally, the image pair and the corresponding camera pose, which can be easily estimated from the input real face, are used to train the 3D generator to synthesize 3D-aware results in the 3D branch. Moreover, we also provide attribute controls (e.g., expression, gender, and age) of 3D cartoonized faces by disentangling features in the intermediate tri-plane spaces. In this manner, users can easily generate their personalized 3D emojis in the similar style as the single style sample they choose.

The main contributions of this work are twofold:

- We proposed a dual-branch network by extensively exploiting the capabilities of 2D and 3D GAN-based generative models, generating stylized 3D avatars without supervision of any 3D data.
- To the best of our knowledge, we are the first to implement one-shot 3D-aware portrait cartoonization for in-the-wild face images, providing a flexible way to generate high-quality personalized 3D emojis with controllable attributes.

2 METHOD

Given an exemplar style image y_e and a portrait image X, we aim to learn the 3D representation of the specific cartoon effect of y_e and generate the cartoonized portrait from any view. Fig. 2 shows the overview of our proposed method. We first perform GAN inversion with a pre-trained encoder and an optimization technique to map the two mages into latent codes in both 2D and 3D domains. Then, we adopt a dual-branch network with 2D and 3D generators trained jointly. The 2D branch acts as a stylization module to provide style guidance for the 3D branch to generate the stylized 3D avatar.

2.1 GAN Inversion

In our model, both the GAN inversions for the 2D branch (StyleGAN) and the 3D generator (EG3D) are needed. A good GAN inverter should be able to balance the trade-off between the quality and editability of the inverted image. More specifically, in the 2D branch, we use an e4e model [Tov et al. 2021] pre-trained on FFHQ to embed the style exemplar y_e and the target input image X to the latent code $w_e \in R^{18\times512}$ and $w_{2d} \in R^{18\times512}$, respectively. Then, we utilize w_e to generate a style dataset to fine-tune the pre-trained StyleGAN to the example cartoon domain via style mixing:

$$w_i = \alpha * w_e + (1-\alpha) * Mapping(z_i), \tag{1}$$

where $Mapping(\cdot)$ denotes the style mapping layers of StyleGAN. Then, w_{2d} is fed into the tuned StyleGAN branch for generating the 2D cartoon portrait to supervise the 3D branch in the next stage.

Given a randomly sampled noise z, the StyleGAN generator of EG3D first maps it to the intermediate latent code w, which is then used to modulate the synthesis network and produce the tri-plane features. Afterwards, the tri-plane features are sampled and decoded to the neural radiance field to generate the final output image. Therefore, for 3D branch, we choose to fit both the intermediate latent code $w \in R^{14\times512}$ and the noise vector n of the synthesis network in the same way as EG3D:

$$w_{3d},\ n = \arg\min_{w_{3d},n} L_{PIPS}(X, G_{EG3D}(w_{3d}, n; \theta)) + \lambda_n L_n(n), \tag{2}$$

where $G_{EG3D}(w_{3d}, n; \theta)$ is the generated image using a fixed pre-trained EG3D synthesis network, decoder, and neural rendering module with weights θ, L_n denotes the noise regularization term, and λ_n is a hyperparameter. Since the in-the-wild image X cannot be fully fitted into the pre-trained feature space, the StyleGAN generator needs to be fine-tuned by minimizing the following loss:

$$\begin{aligned} L_{3d} = {} & L_{PIPS}(X, G_{EG3D}(w_{3d}, n)) + \lambda_{L_2} ||X - G_{EG3D}(w_{3d}, n)||_2 \\ & + \lambda_{ID} L_{ID}(X, G_{EG3D}(w_{3d}, n)). \end{aligned} \tag{3}$$

2.2 Dual Branch Strategy

After adjusting the 3D generator into the in-the-wild image domain, we need to learn the example style representation in the 3D domain. EG3D's [Chan et al. 2022] hybrid explicit-implicit expression of 3D shape based on StyleGAN has been proved effective. Thus it is possible to achieve 3D-aware portrait cartoonization by following the similar idea. In EG3D [Chan et al. 2022], the camera pose condition is introduced to achieve multi-view consistency and disentanglement of some pose-correlated attributes during the training process. Therefore, it is also necessary for us to input the extracted camera pose information into our model to implement cartoonization. Due

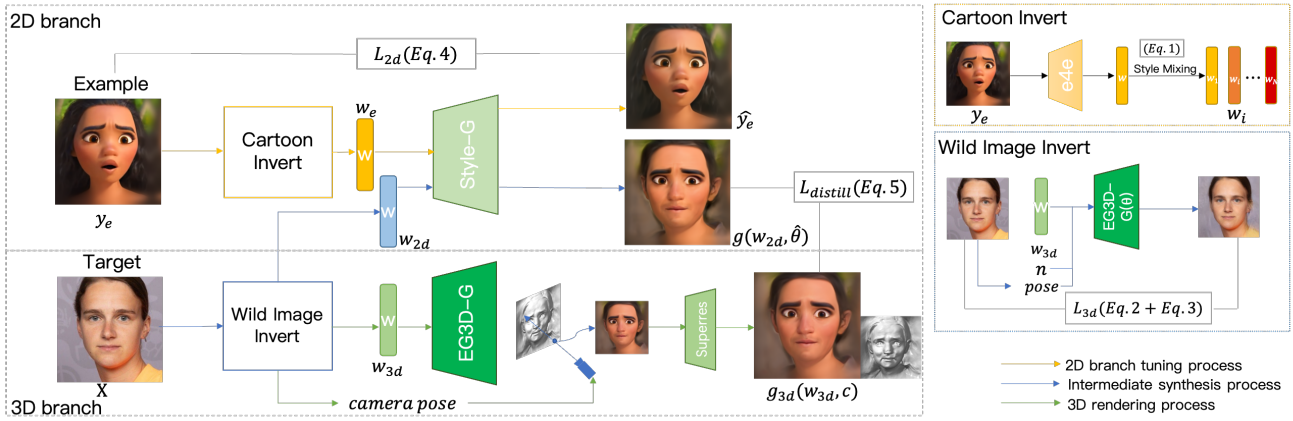

Figure 2: An overview of our model that consists of two parallel but collaborated branches. Given a portrait and a single style image, our model can automatically generate a cartoonized avatar with true-to-life 3D geometry.

to the exaggerated facial features and expressions of cartoon images, large errors often exist when estimating camera poses. Thus we cannot directly use the cartoon dataset to finetune the 3DGAN in this task. Another intuitive solution is to fine-tune the 3DGAN on the paired dataset consisting of the real face images and their corresponding cartoonized ones, whose camera poses can be easily estimated from the real images. However, we experimentally found that the model working in this manner not only has a redundant network architecture but also performs poorly. The major reasons are twofold: 1) the model does not learn to build the domain gap between real faces and cartoon-style images; 2) the model cannot integrate the style information with the 3D-aware representation.

To address the above-mentioned issues, we design a dual-branch network architecture (see Fig. 2). As mentioned in Eq.2, we use w_e to tune a pre-trained StyleGAN into the exemplar style domain, and the pre-trained StyleGAN can be optimized by:

$$\hat{\theta} = \arg\min_{\theta} ||\phi(g(w_e, \theta)), \phi(y_e)||_1 . \tag{4}$$

where $g(w, \theta)$ denotes the synthesis network in StyleGAN and $\phi(\cdot)$ means the mapping via the activation layers of a pre-trained StyleGAN-D. Then, w_{2d} is decoded to a coarsely cartoonized 2D image $g(w_{2d}, \hat{\theta})$, which is served as a supervision for optimizing G_{EG3D} by minimizing:

$$\begin{aligned} L = {} & L_{MSE}(g(w_{2d}, \hat{\theta}), g_{3d}(w_{3d}, c)) + \lambda_1 L_{LPIPS_e}(g_{3d}(w_{3d}, c), y_e) \\ & + \lambda_2 L_{LPIPS_{2d}}(g_{3d}(w_3 d, c), g(w_{2d})). \end{aligned} \tag{5}$$

where $g_{3d}(\cdot)$ represents the Synthesis Network, the Neural Renderer containing a Tri-plane Decoder and a Volume Renderer, and the Super Resolution module in EG3D; $g_{3d}(w_{3d}, c)$ denotes the image with the cartoonized 3D representation for the given target image X with the camera pose condition c. According to the constraints above, the 2D and 3D branches will update their network parameters alternately.

2.3 Attribute Control of 3DGAN

Although EG3D generates images by sampling and rendering on the learned hybrid explicit-implicit tri-plane representation, the tri-plane module and the super-resolution module both operate in the latent space defined by StyleGAN. By inspecting the continuous mapping function and the style synthesis network, it is reasonable to assume that the hybrid 3D representation can also be disentangled via a linear separation boundary, as described in [Karras et al. 2019]. The continuity of the intermediate tri-plane space can be proved experimentally via interpolated style-mixing results.

For a certain binarize facial attribute, such as smiling, gender, eyeglasses, and etc., there should be a hyperplane to classify it. To achieve attribute editing, we should find out the corresponding attribute hyperplane. Inspired by the Linear Separability proposed in StyleGAN, we use an auxiliary linear classifier $f_s(\cdot)$ trained on each attribute, score the latent vector w_{3d} via $f_s(g_{3d}(w_{3d}, c))$, and try to find the boundary $b \in R^{D\times 1}$ via a linear SVM. As mentioned before, the pose condition of our model is determined by an extra camera pose parameter conveyed into the mapping network. Thus, we do not train the pose attribute but directly control the view by the explicit camera condition, which is also the essence of obtaining 3D-aware multi-view consistency.

Attribute editing can be achieved by adding the distance d from the attribute boundary to w_{3d} mentioned above to generate the finale result $g_{3d}((w_{3d} + db^T), c)$, where b^T denotes the normal vector of the boundary b. Fig. 3 shows some results of attribute editing, including smiling, angry, glasses, gender changing, and age changing. In addition, we pre-computed and pre-defined some coefficients, which measure the degrees of attribute changing for expressions, such as smile, surprise, anger, etc., to facilitate the users to generate personalized 3D emojis from their portraits.

3 EXPERIMENTS

3.1 Implementation details

We test our model on 8 NVIDIA Tesla V100 GPUs and use a batch size of 4 per GPU. In our experiments, the style exemplar y_e and

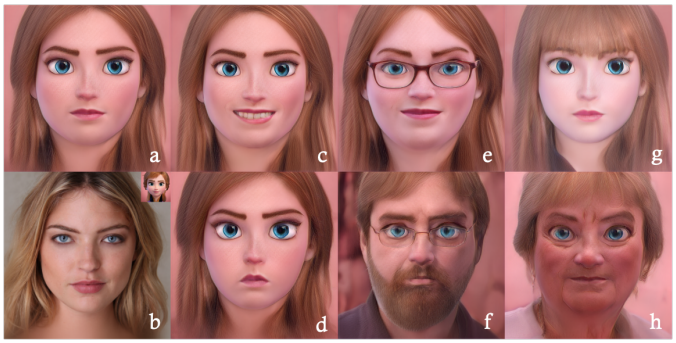

Figure 3: Examples of attribute editing results, (b) shows the target image and the style exemplar. (a) is the cartoonized result. (c-h) are the editing results of smiling, angry, wearing glasses, gender changing, changing young and old.

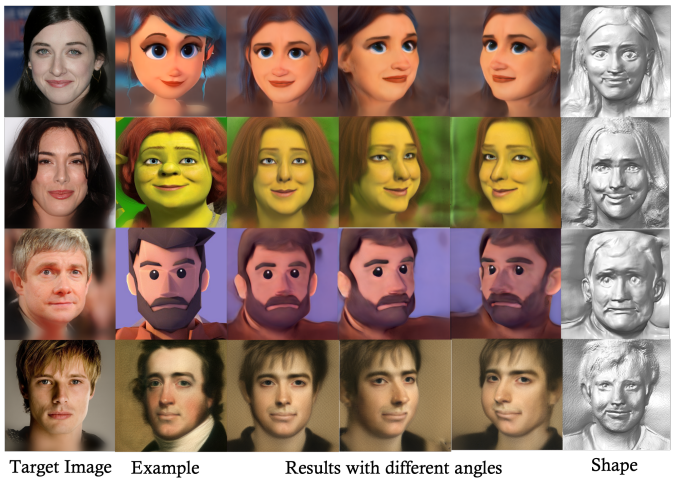

Figure 4: Some stylized 3D avatars generated by our method.

the target image X are both collected from the Internet. We select several art portrait images, e.g. Pixar characters, 3D cartoons, comics, etc., as style references, and the portraits of some football stars as target images. Although our model is designed to be 3D-aware, we find that it also works well for some artistic styles that do not have 3D sense, such as the images in MetFaces [Karras et al. 2020]. All we mentioned generators and encoders are pretrained on FFHQ1024, except for the auxiliary classifiers mentioned in section 2.3, which is on the labeled CelebA via a progressive discriminator. Synthesis results can be found in Fig 4 and our supplementary video. Code and more details can be seen in our project page: https://github.com/41xu/Your3dEmoji.

3.2 Ablation Study

To verify the necessity of our dual-branch supervision, we compare our results with those obtained by directly using the 10k cartoonized FFHQ to tune the 3D branch. Note that the cartoonlization

Figure 5: Ablation study. Synthesis results of EG3D and our method are shown in the first and second rows, respectively

is achieved via the 2D branch of our model. Fig. 5 shows the ablation study results. To demonstrate the capability of our method and the 3D view consistency of the generated results, we also compare our results to the videos that are generated by EG3D and then cartoonized via the 2D branch of our model frame-by-frame. Comparison results can be found in our supplementary video.

4 CONCLUSIONS

This paper proposed a one-shot style transfer method to convert a real-world face image into its corresponding 3D avatar. The key idea is to design a dual-branch network architecture to bridge the domain gap between 2D real faces and 3D cartoon avatars. Experimental results demonstrated that our method could effectively synthesize 3D avatars from 2D images and provide a flexible tool for users to create personalized 3D emojis easily.

ACKNOWLEDGMENTS

This work was supported by Beijing Nova Program of Science and Technology (Grant No.: Z191100001119077) and Project 2020BD020 supported by PKU-Baidu Fund.

REFERENCES

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. 2022. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*. 16123–16133.

Min Jin Chong and David Forsyth. 2021. Jojogan: One shot face stylization. *arXiv preprint arXiv:2112.11641* (2021).

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2021. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985* (2021).

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems* 33 (2020), 12104–12114.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 34 (2021), 852–863.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*. 4401–4410.

Yifang Men, Yuan Yao, Miaomiao Cui, Zhouhui Lian, and Xuansong Xie. 2022. DCT-net: domain-calibrated translation for portrait stylization. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–9.

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

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

Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2022. Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer. In *CVPR*. 7693–7702.