Towards High-Fidelity Text-Guided 3D Face Generation and Manipulation Using only Images
(Supplemental Material)

Cuican Yu[*1], Guansong Lu[*2], Yihan Zeng[*2], Jian Sun[1], Xiaodan Liang[3],
Huibin Li[1], Zongben Xu[1], Songcen Xu[2], Wei Zhang[2], Hang Xu[2†]
[1] Xi'an Jiaotong University [2] Huawei Noah's Ark Lab [3] Sun Yat-sen University
ccy2017@stu.xjtu.edu.cn {luguansong,zengyihan2,xusongcen,wz.zhang}@huawei.com
{jiansun,huibinli,zbxu}@xjtu.edu.cn {xdliang328,chromexbjxh}@gmail.com

In this supplement, we first provide several examples of text-guided 3D object generation methods, and then show additional results to illustrate the effect of the proposed fine-grained text-to-face alignment module. Finally, we introduce the backbone of our method and evaluation metrics.

1. Results of CLIP-Mesh and DreamFusion

To illustrate that the existing methods of generating 3D shapes from the text are difficult to obtain satisfactory results when directly applied to 3D face generation, we show the results of text-guided 3D face generation by CLIP-Mesh [1] and DreamFusion [3]. We adopt the official code of CLIP-Mesh [1], and an unofficial code of DreamFusion [2], since the authors of DreamFusion have not shared the code. As shown in Figure 1, the first three columns in the following figure are the results of ClipMesh according to the top row of input text, and the fourth column is of DreamFusion. As we can see, these methods cannot generate reasonable 3D faces from the text.

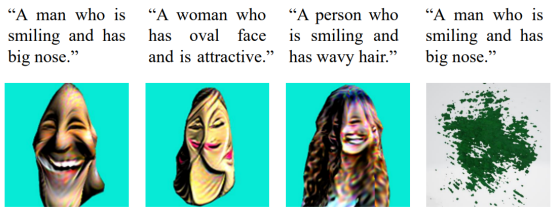

Figure 1. Results of generating 3D face from text by ClipMesh (the first three columns) and DreamFusion (the last columns).

[1]https://github.com/NasirKhalid24/CLIP-Mesh
[2]https://github.com/ashawkey/stable-dreamFusion

2. The Effect of Fine-grained Text-to-Face Alignment

In the fine-grained text-to-face alignment module, similarities between part-level image features and part-level text features is calculated. To illustrate these similarities, we normalize the similarity between each part-level image feature and all part-level text features and fill in the corresponding part. Results presented in Figure 2 illustrate that when a part-level image is semantically close to a textual description, there is a larger similarity between them, suggesting that the feature of the given part will play an important role in the aggregation of part-level image features.

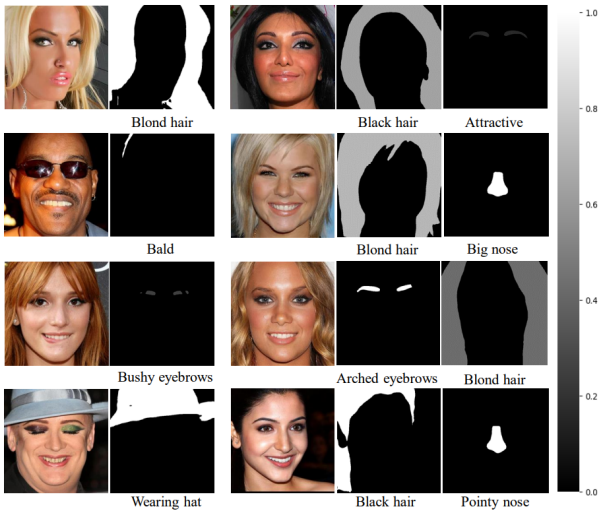

Figure 2. visualization of similarities between part-level image features and part-level text features.

3. Backbone

3.1. Text-conditional Generator

The text-conditional generator consists of a mapping network, a StyleGAN2 generator, a decoder, and neural volume rendering. The mapping network has 8 fully connected layers, in which text embedding, random noise, and camera parameters are concatenated as input. The StyleGAN2 generator outputs a feature map with a size of $256 \times 256 \times 96$, which is then reshaped into three plans, each of shape with a size of $256 \times 256 \times 32$. The decoder is a single fully connected layer of 64 hidden units followed by the softplus activation function. Neural volume rendering [2] adopted in our text-conditional generator is the same as EG3D.

3.2. Text-conditional Discriminator

Following EG3D, we adopt a dual discriminator, which receives high-resolution and low-resolution images at the same time. To condition the discriminator on the input text, we concatenate the output of the dual discriminator of EG3D with the feature of input text and let it go through two additional fully connected layers.

3.3. Fine-Grained Text-to-Face Alignment Module

The fine-grained text-to-face alignment module contains a face parsing model [5], a CLIP text encoder [4], a feature extractor, a linear projection, and a classifier. The face parsing model and the CLIP text encoder are off-the-shelf. The feature extractor is composed of seven residual blocks followed by a fully-connected layer, which outputs a 512-dimension feature for each part of the face image. The linear projection is a single fully connected layer of 512 hidden units to project the part-level text features before calculating the score map between the part-level image features and the part-level text features. The classifier consists of three fully connected layers, in which each layer except the last one is followed by the LeakyReLU activation function, and the last layer outputs 40 binary predictions about facial attributes.

4. Evaluation Metrics

FID. On the Multi-modal CelebA-HQ and CelebAText-HQ datasets, text-to-image generation methods and our TG-3DFace are trained with the training data respectively. Then, text-to-image generation methods generate face images and our TG-3DFace generates 3D faces from texts in the test sets. These 3D faces are rendered into 2D face images with a mean camera pose from the training set. Finally, Frechet Inception Distance (FID) between generated images and images in the test set is calculated respectively on the two datasets by using the code [3].

[3]https://github.com/mseitzer/pytorch-fid

MVIC. Multi-view identity consistency (MVIC) for models trained on Multi-modal CelebA-HQ and CelebAText-HQ are evaluated by measuring cosine similarity of embedding extracted using the 2D face recognition model [4]. For each method on each dataset, we generate 3D faces from texts in the test set and render two views of each face from poses randomly selected from the training dataset. We measure facial identity similarity for each pair and compute the mean score.

[4]https://github.com/timesler/facenet-pytorch

References

[1] Nasir Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-Mesh: Generating textured meshes from text using pretrained image-text models. *ACM Transactions on Graphics, Proc. SIGGRAPH Asia*, 2022. 1

[2] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 2

[3] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

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

[5] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 325–341, 2018. 2