# Towards High-Fidelity Text-Guided 3D Face Generation and Manipulation Using only Images

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

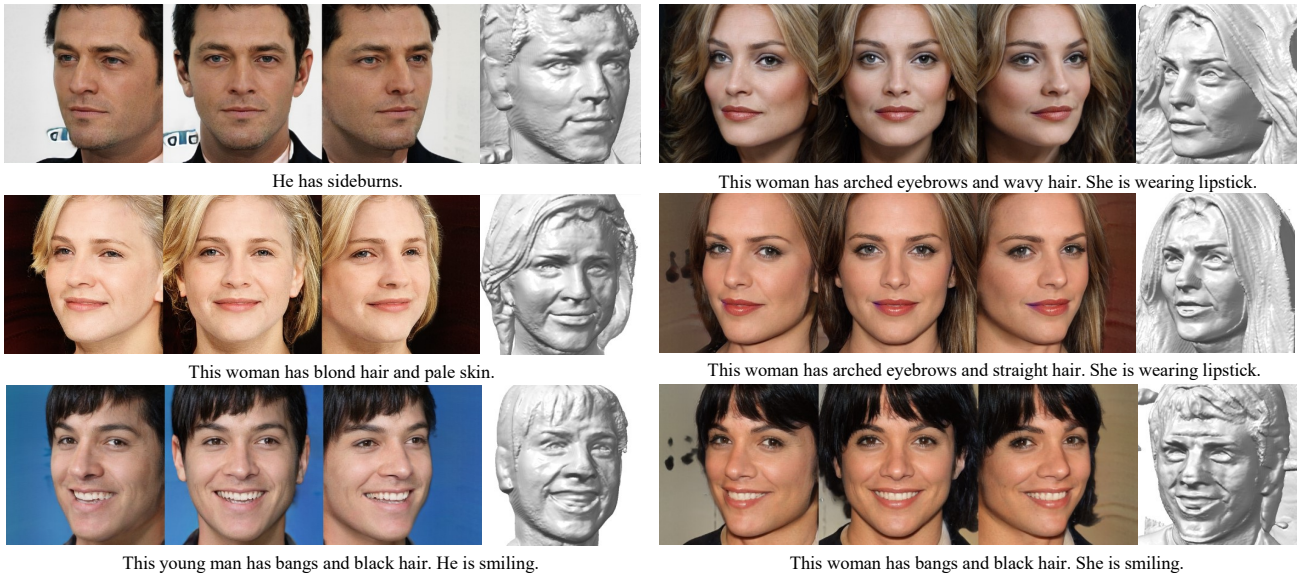


Figure 1. 3D face generations of our TG-3DFace. Given input texts, TG-3DFace can generate high-quality 3D faces with multi-view-consistent rendered face images and detailed 3D face meshes. Notably, fine-grained facial attributes are well controlled by the input texts.

## Abstract

*Generating 3D faces from textual descriptions has a multitude of applications, such as gaming, movie, and robotics. Recent progresses have demonstrated the success of unconditional 3D face generation and text-to-3D shape generation. However, due to the limited text-3D face data pairs, text-driven 3D face generation remains an open problem. In this paper, we propose a text-guided 3D faces generation method, refer as TG-3DFace, for generating realistic 3D faces using text guidance. Specifically, we adopt an unconditional 3D face generation framework and equip it with text conditions, which learns the text-guided 3D face generation with only text-2D face data. On top of that, we propose two text-to-face cross-modal alignment techniques, including the global contrastive learning and the fine-grained alignment module, to facilitate high semantic consistency between generated 3D faces and input texts. Besides, we present directional classifier guidance during the inference process, which encourages creativity for out-of-domain generations. Compared to the existing methods, TG-3DFace creates more realistic and aesthetically pleasing 3D faces, boosting 9% multi-view consistency (MVIC) over Latent3D. The rendered face images generated by TG-3DFace achieve higher FID and CLIP score than text-to-2D face/image generation models, demonstrating our superiority in generating realistic and semantic-consistent textures.*

∗ Equal contribution
† Corresponding author

# 1. Introduction

3D Face generation is a critical technology with diverse applications in various industry scenarios, e.g., movies and games. Recent works have demonstrated the success of 3D face generation with image reconstruction [23, 24] and unconditional generation methods [5, 2, 50, 32, 4]. Despite the photo-realistic 3D face results, the generation process cannot be guided by texts, which has the potential to increase creativity and efficiency. Therefore it is highly demanded to take a step toward text-guided 3D face generation.

Existing methods have been explored to generate 3D shapes and human bodies based on given texts [6, 29, 44, 16, 31], which enables the controllable generation under text guidance. However, it is not feasible to directly apply those generation methods for 3D face generation, owing to two fact:1) The lack of large-scale text-3D face data pairs for model training. 2) The richness of 3D facial attributes that contains much more geometrical details than common 3D objects. Though recent works [3, 1] make attempts to semantically manipulate the shape or texture of 3D faces to boost 3D face generation results, they still lead to results with poor realism and aesthetic appeal such as the loss of hair, which limits the practical applications. Based on the above observation, it requires a rethink of a fine-grained text-driven 3D face generation framework.

To address the above issues, we present a novel fine-grained text-driven 3D face generation framework, named TG-3DFace, to generate high-quality 3D faces that are semantically consistent with the input texts. Specifically, TG-3DFace contains a text-conditioned 3D generation network and two text-to-face cross-modal alignment techniques. Firstly, we adopt the architecture design of EG3D [4], which is an unconditional 3D shape generative adversarial network, and learn 3D shape generation from single-view 2D images. We inject the texture condition into the generator and discriminator networks to enable 3D face generation under the guidance of input texts. Such text-guided 3D face generative model can thus conduct training on text-2D face images instead of text-3D face shapes, enabling to transfer the semantic consistency between texts and 2D face images to guide 3D face generation. Besides, considering the richness of fine-grained facial attributes that increases the difficulty of aligning 3D face and input texts, we design two text-to-face cross-modal alignment techniques, including global text-to-face contrastive learning and fine-grained text-to-face alignment module. The text-to-face contrastive learning aligns the features of the rendered face images with their paired text and maximizes the distance between the unpaired ones in the embedding space, which facilitates global semantic consistency. The fine-grained text-to-face alignment is designed to align the part-level facial features of the rendered face image to the part-level text features, to achieve fine-grained semantic alignment between the texts and the generated 3D faces.

Additionally, we utilize the directional vector in the CLIP embedding space, calculated between the input text and the training style prompt text, as an optimization direction to fine-tune the generator for several steps during inference. In this way, our TG-3DFace can synthesize novel-style face that is never seen during training, such as "a Pixar-style man".

We evaluate our model on the Multi-Modal CelebA-HQ [54], CelebAText-HQ [49] and FFHQ-Text [63] datasets. The experimental results and ablation analysis demonstrate that our method can generate high-quality and semantic-consistent 3D faces given input texts. Besides, our method can be applied to downstream applications including single-view 3D face reconstruction and text-guided 3D face manipulation. In brief, our contributions can be summarized as follows:

- We propose a novel 3D face generation framework, TG-3DFace, which equips the unconditional 3D face generation framework with text conditions to generate 3D faces with the guidance of input texts.
- We propose two text-to-face cross-modal alignment techniques, including global contrastive learning and fine-grained text-to-face alignment mechanism, which boosts the semantic consistency of generations.
- Quantitative and qualitative comparisons confirm that 3D faces generated by our TG-3DFace are more realistic and achieve better semantic consistency with the given textual description.

# 2. Related Work

## 2.1. 3D Face Generation

3DFaceGAN [32] applies the generative adversarial networks (GANs [13]) to represent, generate and translate 3D facial shapes meshes. pi-GAN [5] presents a SIREN network as the generator to represent the implicit radiance field, which conditioned on an input noise. The authors also propose a mapping network with FiLM conditioning and a progressive growing discriminator to achieve high quality results. AvatarMe [23, 24] explores to reconstruct photo-realistic 3D faces from a single "in-the-wild" face image based on the state-of-the-art 3D texture and shape reconstruction method. RigNeRF [2] uses a 3DMM-guided deformable neural radiance field to generate a human portrait trained on a short portrait video. FENeRF [50] proposes to condition a NeRF generator on decoupled shape and texture latent code to learn the semantic and texture representations simultaneously, which helps to generate more accurate 3D geometry. EG3D [4] proposes a tri-plane-based hybrid explicit-implicit 3D representation with a high computational efficiency. They also introduce a dual-discriminator

training strategy to enforce the view-consistency of the final output. In contrast to these works, our method explores text-conditioned 3D face generation with other mechanisms to enforce the semantic consistency between the given text and generated 3D face so that the generated 3D faces can be flexibly controlled by the inputted texts.

### 2.2. Text-to-Image Generation

Given a text description, text-to-image generation aims to generate an image to visualize the context described by the text. There are numbers of works for text-to-image generation along with the generative models, including generative adversarial networks (GANs [13]) [40, 60, 61, 56, 27, 9, 64, 51, 57], auto-regressive model [52, 39, 7, 11, 8, 62, 26, 59] and diffusion model [15, 34, 38, 43, 42]. There are also works focusing on facial image generation. Text2FaceGAN [33] explores to apply the state-of-the-art GAN at the time on text-to-face generation. Stap *et al*. [47] proposes textStyleGAN to generate facial image from text by conditioning the StyleGAN [20] model on text and then manipulate the generated image in the disentangled latent space to make the result semantically more close to the text. SEA-T2F [49] presents a Semantic Embedding and Attention network for multi-caption text-to-face generation. TTF-HD [53], TediGAN [54, 55] and AnyFace [48] propose to align/manipulate the input latent vector of a pretrained StyleGAN model guided by input text to achieve text-to-face generation/manipulation. Recently, PixelFace [35] and OpenFaceAN [36] are proposed to achieve higher performance than StyleGAN-based methods. However, these works only generate single view images and do not consider 3D face generation, while our method outputs high-quality 3D face shapes and multi-view consistent images.

### 2.3. Text-to-3D Shape Generation

Text2shape [6] describes a text-conditioned Wasserstein GAN for generating voxelized 3D objects. CLIP-forge [44] trains a normalizing flow network to generate 3D shape embedding conditioned on CLIP image embedding and uses it to generate 3D shapes conditioned on CLIP text embeddings. Liu *et al*. [29] uses an implicit occupancy representation and proposes a cyclic loss to enforce the consistency between the generated 3D shape and the input text. ShapeCrafter [12] explores recursive text-conditioned 3D shape generation that continuously evolve as phrases are added. However, these works cannot generate realistic 3D objects with high fidelity. Text2mesh [31], CLIP-mesh [22] and Dreamfields [18] use different 3D representations and optimize them with CLIP semantic loss coupled with some regularization terms. DreamFusion [37] instead uses a pretrained 2D text-to-image diffusion model and introduces a novel loss based on probability density distillation to optimize the 3D model. Some works [16, 17, 3] explore 3D avatar generation and animation from text. However, these works cannot generate high-quality 3D faces as 3D faces contain more details than other 3D shapes (like chair, sofa) and human bodies. On the contrary, our method can generate high-quality 3D faces with rich facial attributes described by the given text.

## 3. The Proposed Method

Our goal is to generate realistic 3D faces from text, facing the main challenges of limited text-3D face data and the requirement for semantic alignment between the generated faces and input texts. To address these challenges, we propose TG-3DFace, as shown in Figure 2, which learns to generate 3D faces in condition of input text by using only text-2D face images. In this framework, global text-to-face contrastive learning and fine-grained text-to-face alignment are proposed to improve the semantic consistency between the generated 3D faces and the input texts.

### 3.1. Text-conditional 3D Face Generation

As illustrated in Figure 2, the text embedding $\bar{e} \in \mathbb{R}^{512}$ of input text $s$ is extracted by the CLIP text encoder $E_T$. The sentence embedding $\bar{e}$, camera parameters $p \in \mathbb{R}^{25}$ (including the intrinsic and extrinsic matrices), and a random noise $z \in \mathbb{R}^{512}$ are concatenated and projected into a latent code by the mapping network. This latent code then modulates the convolution kernels of the StyleGAN2 generator, producing a tri-plane representation of the 3D face from which 3D positions can be queried. The features of sampled 3D positions are aggregated and interpreted as a scalar density and a 32-channel feature by the decoder, both of which are then processed by a neural volume renderer [30] to project the 3D feature volume into a 2D face image $\hat{x}$. We define the mapping network, the StyleGAN2 generator, the decoder and the neural volume rendering as our text-conditional generator $G_\theta : \bar{e} \rightarrow \hat{x}$. Subsequently, the text-conditional discriminator $D_\phi$ distinguishes real face images $x$ and the rendered fake images $\hat{x}$ based on the input text embedding $\bar{e}$ and camera parameters $p$.

### 3.2. Text-to-Face Cross-Modal Alignment

#### 3.2.1 Global Text-to-Face Contrastive Learning

In order to generate 3D faces aligned with the input text, we propose global text-to-face contrastive learning to encourage the embeddings $(\bar{e}, \bar{i})$ of paired text and face image close to each other, and unpaired ones $(\bar{e}, \bar{i}^{'})$ away from each other, where embeddings of text and images are extracted by the CLIP text encoder $E_T$ and CLIP image encoder $E_I$ respectively. Global text-to-face contrastive learning encourages the generator to synthesize 3D faces that are semantically consistent with input text. In particular, for each mini-batch, the image $\hat{x}_i$ generated from the input text

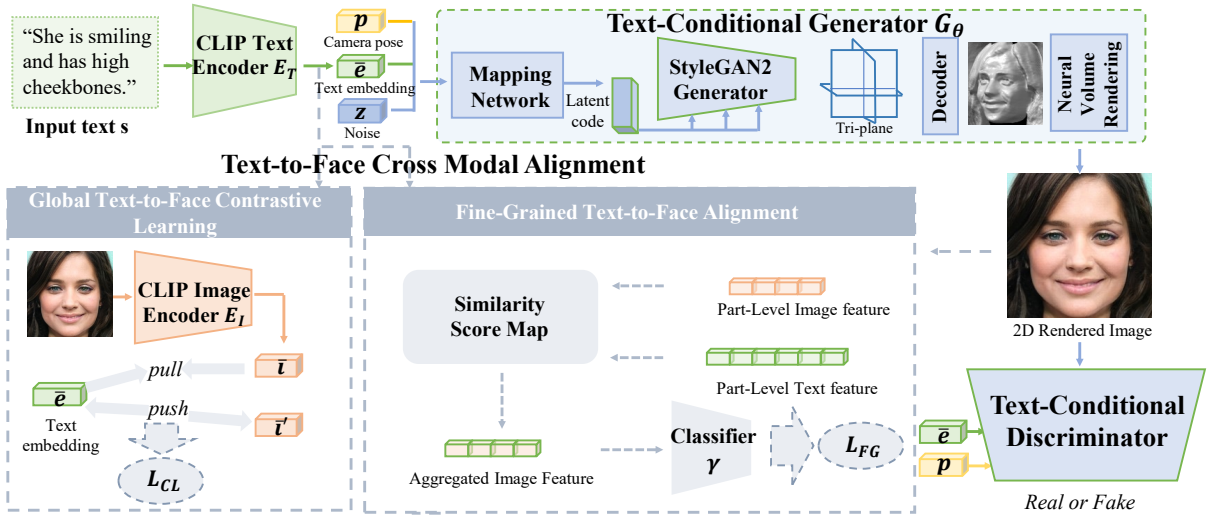


Figure 2. The framework of TG-3DFace. The text-conditional generator synthesizes 3D faces from text embedding $\bar{e}$ extracted by a pre-trained CLIP encoder, then renders them with camera parameters $p$. The text-conditional discriminator distinguishes real face images and the rendered fake face images and is trained adversarially against the generator. The text-to-face contrastive learning module provides a global text-face matching loss for the generator, enhancing the semantic consistency between the input texts and generated 3D shapes. The fine-grained text-to-face matching module helps the generator capture fine-grained semantic content of the input texts more accurately.

$s_i$ is treated as a positive sample of $s_i$, while the other generated images $\hat{x}_j$ are regarded as negative samples. Positive and negative texts to the image $\hat{x_i}$ can be similarly defined.

Formally, the loss function for global text-to-face contrastive learning in a mini-batch is defined as:

$$L_{CL} = \frac{1}{2n}\sum_{i=1}^{n}[L(\hat{x}_i) + L(s_i)], \tag{1}$$

where the loss for input text $s_i$ is defined as follows:

$$L(s_i) = -\frac{1}{n}\log\frac{\exp(E_T(s_i)\cdot E_I(\hat{x}_i)/\tau)}{\sum_{j=1}^{n}\exp(E_T(s_i)\cdot E_I(\hat{x}_j)/\tau)}, \tag{2}$$

where $E_T$ and $E_I$ are the CLIP text encoder and image encoder, $\tau$ is a temperature parameter, $n$ denotes the batch size, and the loss function $L(\hat{x}_i)$ of contrastive learning for the generated image $\hat{x_i}$ can be similarly defined.

### 3.2.2 Fine-grained Text-to-Face Alignment

Until now, it is still challenging for the model to capture fine-grained facial attributes in the input text, as there is no fine-grained supervisions during training. To this end, we explore fine-grained text-to-face alignment training signals for the text-conditional generator.

In fact, facial attributes are mainly displayed across several specific image areas. For example, the image area of the eyes corresponds to the facial attributes like "blue eyes", which means that different face regions have different contributions to an attribute. Inspired by this, we propose to extract part-level image features, and align them with features of a set of pre-defined part-level texts about facial attributes, such as "Black hair", "Mustache", etc. As the fine-grained text-to-face alignment module $C_\varphi$ shown in Figure 3, the rendered face image is first segmented into several parts by an off-the-shelf face parsing algorithm [58]. Then, features of these part-level images are extracted by a feature extractor $\delta$, and a similarity matrix is established between these part-level image features and the fine-grained attribute text description. The part-level image features are then aggregated according to this similarity matrix, as the feature of the input face image. The aggregated feature is computed based on the feature similarity between the fine-grained text and part-level images, thereby focusing on the semantic information of the fine-grained attributes.

Formally, the part-level image features $F = \{f_i\}_{i=1}^{M} \in \mathbb{R}^{M\times d}$ of an image are extracted by a part-level feature extractor $\delta : x_p \rightarrow \mathbb{R}^{512}$, where $x_p$ is a part-level face image in size of $512 \times 512 \times 3$. Part-level text features $H = \{h_j\}_{j=1}^{N} \in \mathbb{R}^{N\times d}$ of pre-defined part-level texts, such as "black hair" and "mustache" are extracted using the CLIP text encoder $E_T$, further projected to a matrix $K = l_K(H) \in \mathbb{R}^{N\times d}$ via learned linear projections $l_K$. The score maps are computed as:

$$W = \text{softmax}(\frac{FK^T}{\sqrt{d}}). \tag{3}$$

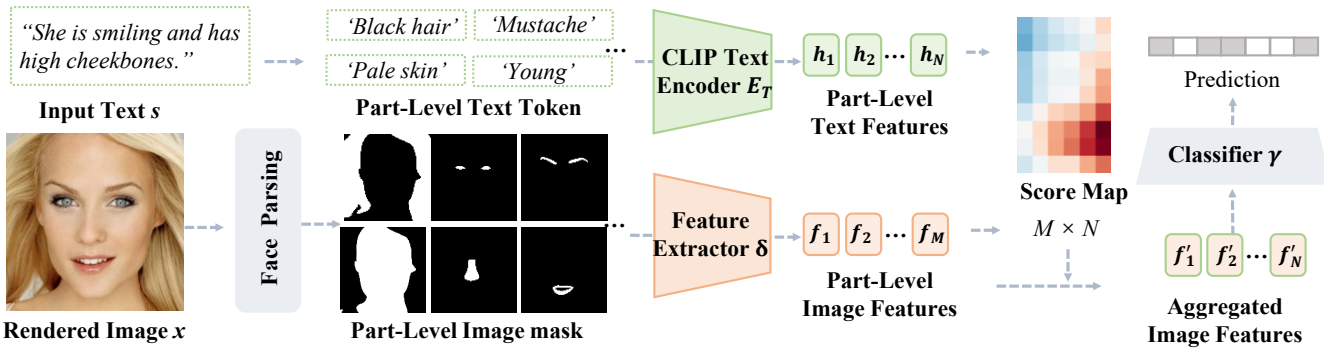


Figure 3. Illustration of fine-grained text-to-face alignment. The rendered image is segmented into several part-level images by face parsing, and then part-level image features are extracted by a learned feature extractor $\delta$. The score map between part-level image features and part-level text features of pre-defined facial attributes text tokens can be used to aggregate the image features. Finally, the aggregated image features is used to predict facial attribute by passing through a learned classifier $\gamma$.

The softmax function is applied to the rows of the scaled similarity matrix, where $d$ is the dimension of part-level image feature and text feature. The aggregated image feature $F^{'} \in \mathbb{R}^{N \times d}$ is then computed as $F^{'} = W^T F$.

Intuitively, the aggregated image feature $F^{'}$ is a weighted average of the part-level image features $F$, where the weights are the score map $W$, which are correlated to the similarity between $F$ and $K$. The $i$-th row of $W$ is a normalized similarity vector between the part-level image feature $f_i$ and all the part-level token features $K$, such that more similar part-level image features have a higher contribution to the combination, whereas less similar part-level image features make little impact. To further enhance the expressiveness of the fine-grained cross-modal alignment, multi-head attention [52] is utilized in parallel before passing the results through a learned classifier $\gamma$ for the final prediction of facial attributes of the input image.

Given the one-hot ground truth label $y \in \mathbb{R}^k$, the facial attribute classification loss can be calculated using the binary cross-entropy loss:

$$L_{FG} = -\frac{1}{k}\sum_{i=1}^{k} y^{(i)} \log(\gamma^{(i)}(F^{'})), \tag{4}$$

where $y^{(i)}$ is the $i$-$th$ ground truth label and $\gamma^{(i)}(F^{'})$ is the $i$-$th$ predicted label of input image $x$ respectively. $k$ is the total number of selected attributes.

### 3.3. Training Loss

During training, the pretrained CLIP text encoder, image encoder, and the face parsing module are frozen, The rest modules in the framework are trained end-to-end. The text-conditional generator $G_\theta$, text-conditional discriminator $D_\phi$ and fine-grained text-to-face alignment module $C_\varphi$ play the following minimax game:

$$\begin{aligned}\min_{\theta,\varphi}\max_{\phi}\frac{1}{n}\sum_{i=1}^{n}\{&\log D_\phi(1 - G_\theta(z_i, s_i, p_i)) + \log D_\phi(x_i) \\ &+ [L_{FG}(G_\theta(z_i, s_i, p_i), y_i) + L_{FG}(x_i, y_i)]\} \\ &- \|\nabla D_\phi(x_i)\|^2 \\ &+ L_{CL},\end{aligned} \tag{5}$$

where $z_i$, $s_i$, $p_i$, $x_i$ and $y_i$ are the $i$-th random noise, input text, camera parameter, real face image and attribute label in the mini-batch sampled from the training dataset, $n$ is the batch size. Specifically, we employ the non-saturating GAN loss function [13], where $G$ is trained to maximize $\log D_\phi(G_\theta(z_i, s_i, p_i))$ rather than $\log D_\phi(1 - G_\theta(z_i, s_i, p_i))$ to provide stronger gradients early in training. The models are trained alternatively from scratch: $D_\phi$ and $C_\varphi$ are firstly trained by one step, and then $G_\theta$ is trained for one step, until converges.

### 3.4. Directional Classifier Guidance

Generally, it is difficult to generate 3D faces from the out-of-domain input text such as "He is a werewolf wearing glasses" since that all our training data are photographs. Inspired by classifier guidance in diffusion models [34, 38, 42], that use an auxiliary discriminative model to guide the sampling process of a pretrained generative model, we utilize the CLIP text encoder and image encoder to design the directional classifier guidance to guide the inference process so as to further improve the text-conditional generator towards generating out-of-domain 3D faces.

Given a target text like "He is a werewolf wearing glasses" denoted as $s^\star$, our generator $G$ generates a 3D face firstly. However, it may be a man wearing glasses as the generator $G$ has never seen werewolf. To perform directional classifier guidance, we clone a copy $G_{frozen}$ of $G$

and freeze it afterwards, and then use $G_{frozen}$ and $G$ to generate a 3D face matching the target text $s^\star$. The directional vector $V_{I_i}$ between the rendered face images from $G$ and $G_{frozen}$ in CLIP space can be obtained with the CLIP image encoder as

$$V_{I_i} = E_I(G(z, s^\star, p_i)) - E_I(G_{frozen}(z, s^\star, p_i)). \quad (6)$$

Similarly, we use a text prompt "Photo" to describe style of training data, noted as $s_o$, and then the directional vector $V_T$ between $s^\star$ and $s_o$ also can be obtained in CLIP space by the CLIP text encoder as

$$V_T = E_T(s^\star) - E_T(s_o). \quad (7)$$

We demand $V_{I_i}$ to be parallel to $V_T$, so that minimize the following directional classifier guidance loss:

$$L_{DCG} = \frac{1}{M}\sum_{i=1}^{M}[1 - \frac{V_{I_i} \cdot V_T}{|V_{I_i}||V_T|}], \quad (8)$$

where $M$ is the number of randomly sampled camera poses in each optimization step. Parameters of the text-conditional generator are updated by the directional classifier guidance loss to synthesize 3D faces matching text $s^\star$. In experiments, we find that the changes in $s^\star$ can result in 3D faces with different styles.

# 4. Experiments

## 4.1. Datasets

We conduct experiments on Multi-Modal CelebA-HQ [54] and CelebAText-HQ [49] datasets to verify the effectiveness of our method for text-guided 3D face generation. The Multi-Modal CelebA-HQ dataset has 30,000 face images, and each one has 10 text descriptions synthesized by facial attributes. The CelebAText-HQ dataset contains 15,010 face images, in which each image has 10 manually annotated text descriptions. All the face images come from the CelebA-HQ [25] dataset, in which each face image has an attribute annotation related to 40 categories, such as "Black hair", "Pale skin", "Young". These attributions are used as pre-defined part-level tokens in our method. In order to learn better 3D face shapes, we added FFHQ [21], a real-world human face dataset without corresponding text description, to the training set. Off-the-shelf facial pose estimators [10, 45] are used to extract approximate camera parameters for each face image in the training set. The FFHQ-Text dataset [63] is a face image dataset with large-scale facial attributes. Since texts in the FFHQ-Text dataset are quite different from those in the Multi-Modal CelebA-HQ dataset, it can be used for cross-dataset experiment.

## 4.2. Metrics

We quantitatively evaluate the generated 3D faces in terms of the quality of their rendered 2D face images, including (1) the multi-view identity consistency (MVIC) by calculating the mean Arcface [46] cosine similarity scores between pairs of face images of the same synthesized 3D face rendered from random camera poses; (2) the reality and diversity of the rendered 2D face images, evaluated by the Frechet Inception Distance (FID) [14]; and (3) the semantic consistency between the input texts and rendered 2D face images, measured by CLIP score. The metric is defined as:

$$\text{CLIPscore}(x, s) = \max(\text{cosine}(E_I(x), E_T(s)) \times 100, 0),$$

which corresponds to the cosine similarity between CLIP embeddings for an image $x$ and a text $s$ respectively in CLIP embedding space, $E_I$ and $E_T$ are CLIP image encoder and CLIP text encoder respectively. Additional evaluation details are introduced in the supplemental materials.

## 4.3. Main Results

In this section, we first qualitatively verify the text-guided 3D face generation ability of our proposed TG-3DFace. Figure 1 shows the input texts and corresponding generated 3D faces, including rendered multi-view face images and 3D face meshes. As we can see, the multi-view face images are consistent with each other and the 3D meshes are detailed, indicating that TG-3DFace is able to generate high-quality 3D faces. Besides, fine-grained facial attributes including "eyebrows", "lipstick", "hair", etc., can be well controlled by the input texts, indicating the text-to-face cross-modal alignment capability of TG-3DFace.

## 4.4. Comparison on Text-to-3D Face Generation

We benchmark the text-guided 3D face generation ability of our proposed TG-3DFace by comparing against a text-to-3D face generation baseline method called Latent3D [3]. Table 1 shows the MVIC scores of Latent3D and TG-3DFace. As we can see, TG-3DFace achieves higher MVIC scores on both datasets, which demonstrates that 3D faces generated by TG-3DFace have better multi-view consistency. As shown in Figure 4, TG-3DFace can generate higher-quality 3D faces with detailed topology (e.g., hairs),

| Methods | Multi-Modal CelebA-HQ | CelebAText-HQ |
|---|---|---|
| Latent3D [3] | 0.87 | 0.85 |
| **TG-3DFace** | **0.95** | **0.93** |

Table 1. Quantitative comparisons on multi-view identity consistency (MVIC) against Latent3D [3] on the Multi-Modal CelebA-HQ and CelebAText-HQ datasets. MVIC is the higher the better.

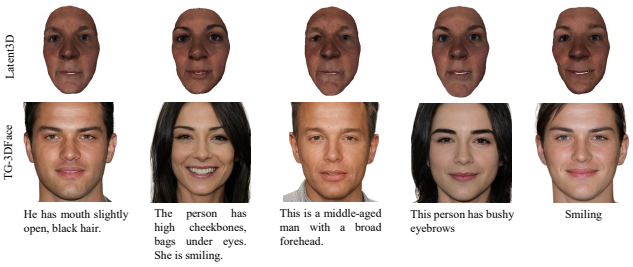


Figure 4. Comparison on text-guided 3D Face generation.

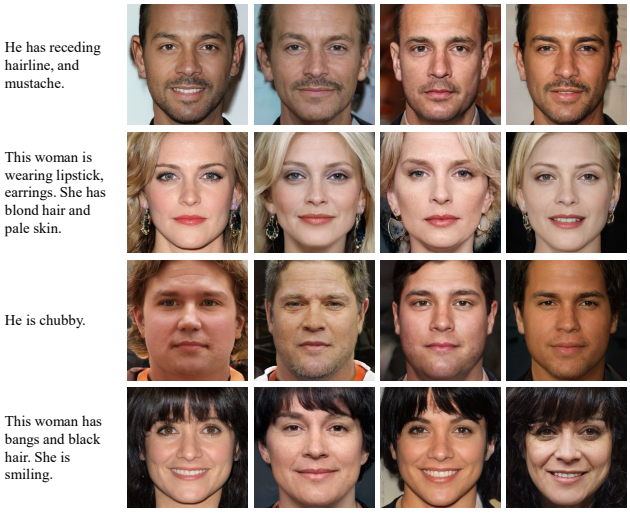


Figure 5. Diverse generation results of TG-3DFace. In each row, we show input text and generated results of different input noises.

as well as realistic facial texture. The generated 3D faces of TG-3DFace with different input texts are also more diverse compared with Latent3D. To further verify the diversity of our results, we show results from the same input text but different input noises in Figure 5. As we can see, given the same input text, TG-3DFace can generate diverse 3D faces according to different input noises.

### 4.5. Comparison on Texture Quality

In this section, we further benchmark the texture quality of generated 3D faces achieved by TG-3DFace and previous start-of-the-art text-to-2D face/image generation methods, including SEA-T2F [49], ControlGAN [27], AttnGAN [56] and AnyFace [48]. As shown in Table 2 and Table 3, our method achieves better FID and CLIP score on both datasets, indicating the texture of generated 3D faces of TG-3DFace are of higher fidelity and semantic consistency.

### 4.6. User Study

We also employ a user study on the CelebAText-HQ dataset as an additional evaluation. 28 images generated from texts by different methods are randomly selected and 38 graduate students are invited to rank these images with

| Methods | FID ↓ | CLIP Score ↑ |
|---|---|---|
| SEA-T2F [49] | 93.85 | 20.81 |
| ControlGAN [27] | 74.59 | 21.38 |
| AttnGAN [56] | 51.69 | 21.52 |
| AnyFace [48] | 50.56 | - |
| **TG-3DFace** | **39.02** | **22.72** |

Table 2. Quantitative comparisons of different methods on the Multi-Modal CelebA-HQ dataset, where ↓ means the lower the better while ↑ means the opposite.

| Methods | FID ↓ | CLIP Score ↑ |
|---|---|---|
| SEA-T2F [49] | 125.32 | 19.06 |
| ControlGAN [27] | 78.01 | 20.70 |
| AttnGAN [56] | 70.59 | 20.17 |
| AnyFace [48] | 56.75 | - |
| **TG-3DFace** | **52.21** | **21.03** |

Table 3. Quantitative comparisons of different methods on the CelebAText-HQ dataset, where ↓ means the lower the better while ↑ means the opposite.

| Methods | Avg-rank on Fidelity ↓ | Avg-rank on Semantic Consistency ↓ |
|---|---|---|
| SEA-T2F [49] | 3.19 | 4.00 |
| ControlGAN [27] | 2.92 | 2.61 |
| AttnGAN [56] | 2.15 | 2.11 |
| **TG-3DFace** | **1.02** | **1.28** |

Table 4. User study. Users show a significant preference for our TG-3DFace over SEA-T2F, ControlGAN and AttnGAN for fidelity and semantic consistency.

| Methods | Avg-rank on Fidelity ↓ | Avg-rank on Semantic Consistency ↓ |
|---|---|---|
| SEA-T2F [49] | 3.49 | 3.18 |
| ControlGAN [27] | 3.37 | 3.16 |
| AttnGAN [56] | 2.13 | 2.36 |
| **TG-3DFace** | **1.00** | **1.30** |

Table 5. User study with out-of-distribution texts. Users show a significant preference for our TG-3DFace over SEA-T2F, ControlGAN and AttnGAN for fidelity and semantic consistency.

the questions "*Are these images real*" and "*Are these images achieve the attributes specified in the text*" (rank 1 is the best). Ranking results for each method are averaged, defined as Avg-rank on Fidelity and Avg-rank on Semantic Consistency. As shown in Table 4, our model achieves better results, indicating our generated textures of 3D faces are of higher fidelity and semantic consistency with input texts.

To validate our TG-3DFace with out-of-distribution texts, we train our model and compared models on the Multi-Modal CelebA-HQ dataset, and test them using texts

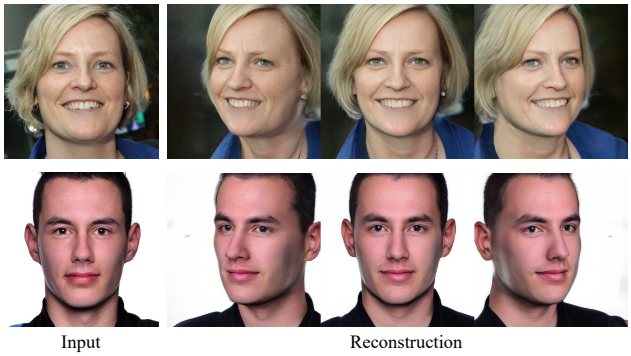


Figure 6. Single-view 3D reconstruction results of TG-3DFace.

| Methods | FID ↓ | CLIP Score ↑ |
|---|---|---|
| w/o $L_{CL}$ | 52.95 | 21.50 |
| w/o $L_{FG}$ | 50.11 | 21.86 |
| w/o $L^*_{FG}$ | 52.57 | 22.03 |
| Full model | **39.02** | **22.72** |

Table 6. Ablation study on Multi-Modal CelebA-HQ dataset, where ↓ means the lower the better, ↑ means the opposite.

| Methods | FID ↓ | CLIP Score ↑ |
|---|---|---|
| w/o $L_{CL}$ | 57.51 | 20.61 |
| w/o $L_{FG}$ | 60.57 | 19.50 |
| w/o $L^*_{FG}$ | 56.75 | 20.73 |
| Full model | **52.21** | **21.03** |

Table 7. Ablation study on CelebAText-HQ dataset, where ↓ means the lower the better and ↑ means the higher the better.

in the FFHQ-Text dataset. We randomly select 28 images generated from texts by different models respectively and invite 30 graduate students to rank them. As listed in Table 5, our model achieves better results on user study with out-of-distribution texts.

### 4.7. Ablation Studies

To analyze the effectiveness of the proposed global text-to-face contrastive learning and fine-grained cross-modal alignment, we conduct ablation studies by removing one of them each time and report the quantitative results on the Multi-Modal CelebA-HQ dataset and CelebaTextHQ datasets. As the results listed in Table 6 and Table 7, omitting $L_{CL}$ and $L_{FG}$ adversely affect the FID and CLIP score. Specifically, the FID increases from 39.02 to 52.95/50.11, and the CLIP score decreases from 22.72 to 21.50/21.86 on the Multi-Modal-celebA-HQ dataset. This highlights the importance of these two modules in improving the quality of generated 3D faces and enhancing the semantic matching between the generated 3D faces and the input texts.

To further emphasize the significance of the fine-grained text-to-face alignment module, we retain its network and loss while removing the operations of face parsing and part-level feature aggregation, and extract features from the full image to predict its attributes instead, denoted as "w/o $L^*_{FG}$". The comparison between "w/o $L^*_{FG}$" and the full model demonstrates the importance of our carefully designed fine-grained cross-modal alignment, ruling out the possibility of it being replaced by an attribute classifier on the full image.


Figure 7. Comparison on text-guided 3D face manipulation.

### 4.8. Applications

**Single-view 3D Face Reconstruction.** Figure 6 illustrates the utilization of our learned latent space for single-view 3D reconstruction via pivotal tuning inversion (PTI) [41]. The 3D prior over text-guided 3D face generation enables impressive single-view geometry recovery. Building on this ability, we can further edit the reconstruction result using directional classifier guidance, which may serve as a promising area for future research.

**Text-guided 3D Face Manipulation.** The proposed directional classifier guidance enables our method to create many interesting results in diverse styles based on our learned text-guided 3D face generative model. Given a text that differs significantly from the texts in training data during inference time, the proposed directional classifier guidance can be used to optimize the generator for a few minutes

to synthesize 3D faces with styles outside the training set. Figure 7 compares 3D face editing results achieved by different text-driven 3D texture manipulation methods, including Latent3D [3], Clip-Matrix [19], FlameTex[1], Text2Mesh [31], ClipFace [1] and TG-3DFace. As we can see, Clip-Matrix, FlameTex, and Text2Mesh cannot get the correct texture according to the text. Latent3D cannot capture finer-grained localization of manipulations, such as changing the color of eyes, and it cannot generate textures in various styles, such as Pixar. Although ClipFace can generate the corresponding texture of 3D faces according to the input texts, they can not handle accessories like headwear, or eyewear, due to the use of the FLAME [28] model, which does not capture accessories or complex hair. Our approach yields consistently high-quality textures for various prompts, in comparison to these baselines. In general, our generator enables high-quality editing, and the style will be more obvious in texture.

### 4.9. Parameters and Runtime

Table 8 compares parameters and inference time running on a single NVIDIA Tesla V100 GPU between our TG-3DFace and several existing text-guided 3D face or object generation methods. We can see that when the model parameters of TG-3DFace are not large, the inference time to generate a 3D face is only 0.05 seconds, and the manipulation time is only 1.5 minutes.

| Methods | Total Params | Trainable Params | Inference Time |
|---|---|---|---|
| Latent3D [3] | 661 M | - | 6 min |
| Clip-Matrix [19] | 154 M | 2.6 M | 30 min |
| Text2Mesh [31] | 151 M | 659 K | 25 min |
| TG-3DFace (generation) | 240 M | 76 M | 0.05 s |
| TG-3DFace (manipulation) | 190 M | 39 M | 1.5 min |

Table 8. Comparison of total parameters, trainable paramateres, and inference time per sample.

## 5. Implementation Details

We train our model with a batch size of 32, and use a discriminator learning rate of 0.002 and a generator learning rate of 0.0025. Similar to EG3D [4], we blur images when they enter the discriminator, gradually reducing the amount of blur of the first 200 K images, and we train our model without style mixing regularization. According to EG3D, low neural rendering resolutions (e.g., 64) enable faster speed of training and inference, while higher neural rendering resolutions (e.g., 128) facilitate more detailed shapes and more view-consistent 3D renderings. Following EG3D, the neural rendering resolution is gradually increased from $64^2$ to $128^2$ over 1 million images during training. The total training time of our model on 8 NVIDIA Tesla A100 GPUs is 48 hours. When the directional classifier guidance is used in the inference phase, the generator is optimized for 100 iterations at a learning rate of 0.002.

[1] https://github.com/HavenFeng/photometric_optimization

## 6. Conclusion

In this paper, we propose a novel method named TG-3DFace for generating high-quality 3D faces with multi-view consistent and photo-realistic rendered face images. Specifically, a text-conditional 3D face GAN enables the model can be trained from text-face images rather than the supervision of 3D faces. Global text-to-face contrastive learning and fine-grained text-to-face alignment modules are proposed to improve the semantic consistency between the generated 3D faces and input texts. Furthermore, we extend our model to synthesize out-of-domain 3D faces by introducing directional classifier guidance. Extensive experimental studies manifest the effectiveness of our method.

## 7. Limitation and Future Work

First, our method cannot infer identity information from textual descriptions, such as "Toms Bond". Second, the 3D faces generated by our method are sometimes asymmetry, such as wearing only one earring. Third, the race of the generated faces is similar to that of training data. We will consider improving the quality of generated shape, and will expand the races of training images so that the resulting faces are not limited to a single race.

## 8. Acknowledgement

The work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant (No. 61976173, 12125104, U20B2075), Shaanxi Fundamental Science Research Project for Mathematics and Physics (Grant No. 22JSY011). We thank MindSpore for the partial support of this work, which is a new deep learning computing framework[2].

## References

[1] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. ClipFace: Text-guided editing of textured 3D morphable models. In *SIGGRAPH '23 Conference Proceedings*, 2023.

[2] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. RigNeRF: Fully controllable neural 3D portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20364–20373, 2022.

[2] https://www.mindspore.cn

[3] Zehranaz Canfes, M Furkan Atasoy, Alara Dirik, and Pinar Yanardag. Text and image guided 3D avatar generation and manipulation. *arXiv preprint arXiv:2202.06079*, 2022.

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

[5] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5799–5809, 2021.

[6] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian conference on computer vision*, pages 100–116, 2018.

[7] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. CogView: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[8] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. CogView2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

[9] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017.

[10] Paul Ekman and Wallace V. Friesen. Facial action coding system: A technique for the measurement of facial movement. 1978.

[11] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. ImageBART: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34:3518–3532, 2021.

[12] Rao Fu, Xiao Zhan, Yiwen Chen, Daniel Ritchie, and Srinath Sridhar. ShapeCrafter: A recursive text-conditioned 3D shape generation model. *arXiv preprint arXiv:2207.09446*, 2022.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[16] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. AvatarCLIP: Zero-shot text-driven generation and animation of 3D avatars. *ACM Transactions on Graphics*, 41(4):1–19, 2022.

[17] Li Hu, Jinwei Qi, Bang Zhang, Pan Pan, and Yinghui Xu. Text-driven 3D avatar animation with emotional and expressive behaviors. In *Proceedings of the ACM International Conference on Multimedia*, pages 2816–2818, 2021.

[18] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022.

[19] Nikolay Jetchev. ClipMatrix: Text-controlled creation of 3D textured meshes. *arXiv preprint arXiv:2109.12922*, 2021.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[22] 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*, 2022.

[23] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. AvatarMe: Realistically renderable 3D facial reconstruction ··in-the-wild". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2020.

[24] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. AvatarMe++: Facial shape and BRDF inference with photorealistic rendering-aware GANs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2021.

[25] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5549–5558, 2020.

[26] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.

[27] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Controllable text-to-image generation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2065–2075, 2019.

[28] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194–1, 2017.

[29] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. Towards implicit text-guided 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906, 2022.

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

[31] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2Mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022.

[32] Stylianos Moschoglou, Stylianos Ploumpis, Mihalis A Nicolaou, Athanasios Papaioannou, and Stefanos Zafeiriou. 3DFaceGAN: Adversarial nets for 3D face representation, generation, and translation. *International Journal of Computer Vision*, 128(10):2534–2551, 2020.

[33] Osaid Rehman Nasir, Shailesh Kumar Jha, Manraj Singh Grover, Yi Yu, Ajit Kumar, and Rajiv Ratn Shah. Text2FaceGAN: Face generation from fine grained textual descriptions. In *International Conference on Multimedia Big Data*, pages 58–67, 2019.

[34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[35] Jun Peng, Xiaoxiong Du, Yiyi Zhou, Jing He, Yunhang Shen, Xiaoshuai Sun, and Rongrong Ji. Learning dynamic prior knowledge for text-to-face pixel synthesis. In *Proceedings of the ACM International Conference on Multimedia*, pages 5132–5141, 2022.

[36] Jun Peng, Han Pan, Yiyi Zhou, Jing He, Xiaoshuai Sun, Yan Wang, Yongjian Wu, and Rongrong Ji. Towards open-ended text-to-face generation, combination and manipulation. In *Proceedings of the ACM International Conference on Multimedia*, pages 5045–5054, 2022.

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

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831, 2021.

[40] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069, 2016.

[41] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics*, 42(1):1–13, 2022.

[42] 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.

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022.

[45] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[46] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–823, 2015.

[47] David Stap, Maurits Bleeker, Sarah Ibrahimi, and Maartje ter Hoeve. Conditional image generation and manipulation for user-specified content. *arXiv preprint arXiv:2005.04909*, 2020.

[48] Jianxin Sun, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. AnyFace: Free-style text-to-face synthesis and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18687–18696, 2022.

[49] Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, and Zhenan Sun. Multi-caption text-to-face synthesis: Dataset and algorithm. In *Proceedings of the ACM International Conference on Multimedia*, pages 2290–2298, 2021.

[50] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. FENeRF: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022.

[51] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[53] Tianren Wang, Teng Zhang, and Brian Lovell. Faces a la carte: Text-to-face generation via attribute disentanglement. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3380–3388, 2021.

[54] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. TediGAN: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2256–2265, 2021.

[55] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Towards open-world text-guided face image generation and manipulation. *arXiv preprint arXiv:2104.08910*, 2021.

[56] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[57] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021.

[58] 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.

[59] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[60] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[61] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

[62] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. M6-UFC: Unifying multi-modal controls for conditional image synthesis. *arXiv preprint arXiv:2105.14211*, 2021.

[63] Yutong Zhou. Generative adversarial network for text-to-face synthesis and manipulation. In *Proceedings of the ACM International Conference on Multimedia*, pages 2940–2944, 2021.

[64] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019.