Computational Visual Media
https://doi.org/10.1007/s41095-024-0422-4

Research Article

Multi3D: 3D-aware multimodal image synthesis

Wenyang Zhou[1], **Lu Yuan**[2], **and Taijiang Mu**[1] (✉)

Abstract 3D-aware image synthesis has attained high quality and robust 3D consistency. Existing 3D controllable generative models are designed to synthesize 3D-aware images through a single modality, such as 2D segmentation or sketches, but lack the ability to finely control generated content, such as texture and age. In pursuit of enhancing user-guided controllability, we propose Multi3D, a 3D-aware controllable image synthesis model that supports multi-modal input. Our model can govern the geometry of the generated image using a 2D label map, such as a segmentation or sketch map, while concurrently regulating the appearance of the generated image through a textual description. To demonstrate the effectiveness of our method, we have conducted experiments on multiple datasets, including CelebAMask-HQ, AFHQ-cat, and shapenet-car. Qualitative and quantitative evaluations show that our method outperforms existing state-of-the-art methods.

Keywords generate adversarial networks (GANs); neural radiation field (NeRF); 3D-aware image synthesis; controllable generation

1 Introduction

The generation of high quality realistic images has wide application in artistic creation. Generative adversarial networks (GANs) [1] are instrumental in learning the mapping from a Gaussian distribution to the distribution of real images through the joint adversarial training of both a generator and a discriminator. Subsequent to the introduction of GANs, numerous studies [2–4] have made noteworthy advances in enhancing image quality and resolution. To exert control over generated images, there has been a surge of research in the field of image-to-image transformation. Various methods [5–7] have been developed to convert a two-dimensional label image into a tangible image that adheres to the semantic directions conveyed by the label.

The evolution of neural radiance fields (NeRFs) has facilitated the bridging of the two-dimensional image and the three-dimensional world, enabling the seamless transfer of gradients between them. Researchers have directed their focus toward 3D GANs, and several methods [8–10] have emerged, showcasing the ability to generate high-quality, three-dimensionally consistent objects. Leveraging the inherent differentiability of NeRFs, these approaches can be entirely trained using two-dimensional images, obviating the necessity for dedicated three-dimensional datasets. Further advances in this field are exemplified by various methods [11–14] that specifically address the utilization of two-dimensional labels for the generation and manipulation of three-dimensional objects. This dedicated research aims to refine and enhance synthesis and editing capabilities, offering valuable insights into the potential of leveraging two-dimensional information for comprehensive three-dimensional modeling.

Nevertheless, existing work predominantly addresses the task of control through a single modality, such as a segmentation or a sketch map. These methods control 3D-aware image synthesis through 2D label maps but are limited in their ability to finely control the content of the generated images, such as age, gender, color, etc. This paper, in contrast, undertakes the more intricate challenge of orchestrating the generation of 3D objects by harnessing multiple input modalities. As illustrated in Fig. 1, our objective is to govern the geometric shape of the generated three-dimensional object

1 BNRist, Tsinghua University, Beijing 100084, China. E-mail: W. Zhou, zhouwy19@mails.tsinghua.edu.cn; T. Mu, taijiang@tsinghua.edu.cn (✉).

2 Computer Science Department, Stanford University, California 94305, USA. E-mail: luyuan@stanford.edu.

Manuscript received: 2023-12-09; accepted: 2024-02-29

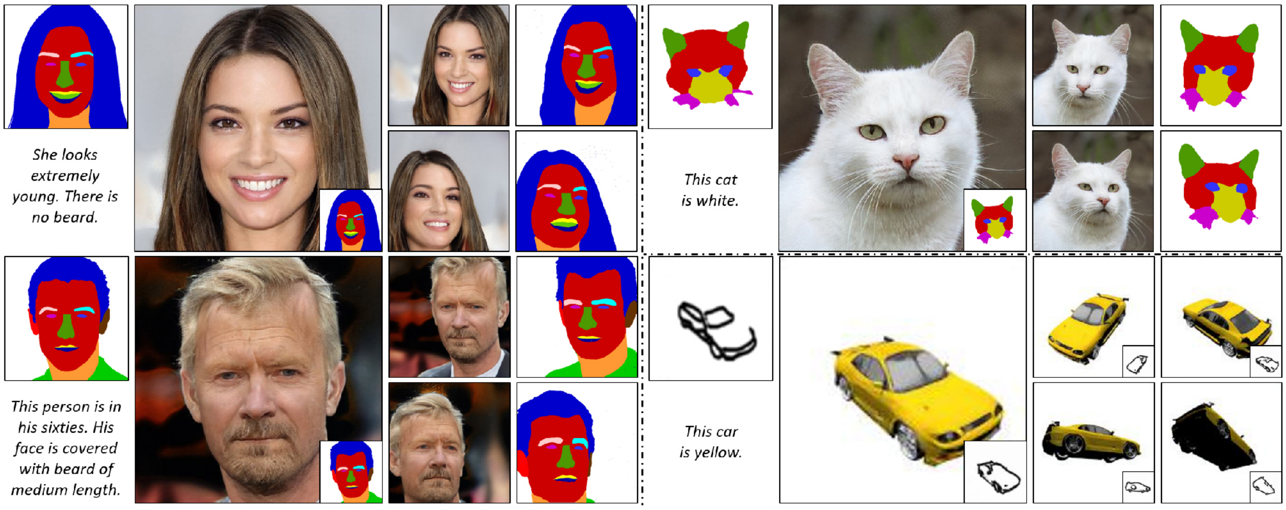
She looks extremely young. There is no beard.
This cat is white.
This person is in his sixties. His face is covered with beard of medium length.
This car is yellow.

Fig. 1 Given 2D segmentations or sketches and descriptive text as input, Multi3D can generate high-quality 3D-aware images that closely meet the input conditions. To demonstrate the effectiveness of our approach, we present generation results on multiple datasets, including CelebAMask-HQ, AFHQ-cat, and shapenet-car.

through a two-dimensional label map. Subsequently, we articulate the appearance characteristics of the three-dimensional object through a descriptive sentence, thereby achieving synthesis of a high-quality, three-dimensionally consistent object in agreement with multiple conditional inputs.

To achieve this goal, several challenges must be addressed: how to (i) inject multiple conditional inputs into the generation process, (ii) decouple multiple conditions to ensure their independence, and (iii) ensure that the generated three-dimensional objects agree with multiple conditional inputs. To tackle the first challenge, we introduce a multimodal condition encoder designed to map various conditions to distinct control vectors. For the second challenge, we implement a conditional cross-training mechanism, enhancing the model's robustness and decoupling capabilities. To address the third challenge, we propose an image text alignment adaptive CLIP loss and a label reconstruction loss, ensuring the alignment of the generated three-dimensional object with multiple conditional inputs simultaneously.

Qualitative and quantitative experiments demonstrate that our method outperforms state-of-the-art approaches in the domain of 3D-aware multimodal image synthesis. The primary contributions of this paper are in summary:

- We pose a new and more challenging task in 3D-aware image synthesis: controllable generation under multiple input conditions.
- We propose Multi3D, a 3D generative model capable of simultaneously controlling the generation process of 3D-aware images using 2D label maps and text. We propose two training techniques, including *adaptive CLIP fine-tuning* (*ACF*) and a *conditional crossover strategy* (*CCS*) , which improve the quality of generated images and alignment with the input conditions.
- Our method achieves better generation and control effects than state-of-the-art 3D-aware conditional image synthesis methods, for diverse datasets, including CelebAMask-HQ [15], AFHQ-cat [16], and shapenet-car [17]. Both qualitative and quantitative evaluations demonstrate the effectiveness of our method for 3D-aware multimodal image synthesis.

2 Related work

Our work is closely related to neural implicit representations and 3D-aware image synthesis.

2.1 Neural implicit representations

One of the seminal works in neural implicit representations is the *neural radiance field* (NeRF), a model that conceptualizes a three-dimensional scene as a neural network. NeRF models take three-dimensional position encoding as input, and predict density and color within the network, employing voxel rendering operations for image synthesis. NeRFs are widely employed for tasks such as 3D reconstruction and the synthesis of novel perspectives. Subsequent methods [8, 19–21] have advanced the field by introducing novel three-dimensional representations, specifically designed to accelerate training and inference processes.

EG3D [8] introduced the tri-planes expression format and implemented it for 3D generation, yielding results characterized by high quality and three-dimensional consistency. It is an unconditional generative model that transforms sampled random noise into a three-dimensional object. Our aim is to enhance the controllability of this generation process through multi-modal conditional forms. To achieve this objective, we have developed a multi-modal conditional encoder and designed a series of loss functions and training strategies to ensure the generated results are consistent with the specified input conditions.

2.2 3D-aware image synthesis

Generative adversarial networks (GANs) [1], introduced in 2014, operate through the adversarial training of a generator and discriminator. Initially applied to generate low-resolution images like those in MNIST and CIFAR-10, subsequent methodologies [2–4] have been dedicated to enhancing the quality and resolution of generated images, scaling up to resolutions of 512^2 or even 1024^2. To enable control over generated images, conditional GANs (cGANs) integrate category information into the generation process, facilitating control over the category of the generated images. Various methods [5–7, 22–38] employ two-dimensional label maps, such as segmentations or sketch maps, to control the generation and editing of images.

The emergence of the neural radiance field has played a pivotal role in advancing the domain of 3D generative adversarial networks. Several methods [9, 10, 39–42] generate high-quality 3D objects by supervising rendered images from different viewpoints. EG3D [8] introduces the tri-planes representation, enhancing both geometric quality and three-dimensional consistency of generated objects. At the same time, many 3D downstream tasks have also undergone rapid development, such as segmentation map-based editing [11–13, 43, 44], sketch-based editing [14], relighting [45, 46], and animation [46, 47]. Several methods [11, 13] adopt a dual-phase strategy where an initial generator is trained to produce both images and semantic label maps. Subsequently, an optimization process on latent vectors refines the generated output according to input label maps. Meanwhile, some methods [33, 44, 48, 49] focus on single-modal 3D image synthesis. Pix2NeRF [48] proposes a NeRF-based single-modality generation model for transforming images into 3D models; however, its impact is constrained by limitations in both quality and three-dimensional consistency. pix2pix3D [49] proposes an end-to-end network and training strategy geared towards the direct conversion of semantic label maps into three-dimensional objects. Existing methods concentrate on single modalities and geometries, which limits the means to control the generation process. Our approach diverges by placing emphasis on novel forms of multimodal control. Specifically, we address hybrid controls that encompass semantic segmentation maps, sketch maps, and textual descriptions. This distinctive focus allows for a more comprehensive and versatile means of governing the generation process, accommodating a diverse set of user inputs across different modalities.

3 Method

3.1 Architecture

Given a 2D segmentation map or sketch, along with descriptive text, Multi3D aims to generate 3D-aware images that satisfy the combined input conditions. Figure 2 shows the architecture of Multi3D. In Section 3.2, we first introduce the three-dimensional representation and generation process of EG3D. In Section 3.3, we introduce the proposed model architecture and training strategy of Multi3D. Our *multi-modal conditional encoders* module is designed to encode multi-modal input. Our adaptive CLIP fine-tuning and conditional crossover strategy respectively improve the quality of generated objects and the alignment of generated objects with text. Subsequently, we elaborate on the loss function used for training Multi3D in Section 3.4, including label reconstruction loss, adaptive CLIP loss, and cross-view consistency loss.

3.2 Preliminaries

Since we use tri-planes proposed by EG3D [8] as a 3D representation in our model, we first introduce the generation process of EG3D, which is initiated by

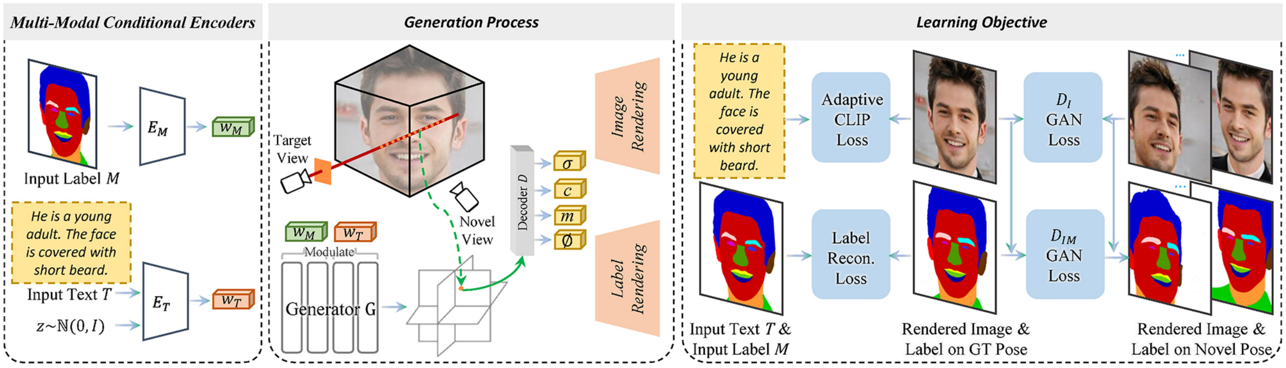

Fig. 2 Framework. Multi3D takes a 2D input label M, a sentence of input text T, and random noise z as inputs. We use our *multi-modal conditional encoders* module to encode them into W latent codes and inject them into the first and last seven layers of the generator G. We adopt tri-planes [8] as our three-dimensional representation. The decoder D is responsible for predicting the density σ, color c, label m, and features ϕ of the three-dimensional sampling points. Low-resolution images and labels are obtained through volume rendering [18]. Subsequently, high-resolution images and labels are obtained through two super-resolution networks. Adaptive CLIP loss and label reconstruction loss supervise the alignment of the generated objects with input labels and text respectively. Two discriminators D_I, D_{IM} are introduced to supervise the quality and alignment of generated images and labels.

sampling random noise $z \in \mathbb{R}^{1\times512}$ and concatenating it with the camera pose embedding $p \in \mathbb{R}^{1\times512}$. These values are then input into several layers of fully connected networks, resulting in the generation of latent codes $w \in \mathbb{R}^{1\times512}$. Subsequently, these latent codes w are transmitted to the generator backbone of StyleGAN2 [4], producing tri-planes $F \in \mathbb{R}^{256\times256\times96}$. The tri-planes F are further partitioned into three orthogonal feature planes $F_{xy}, F_{yz}, F_{xz} \in \mathbb{R}^{256\times256\times32}$. Each sampled three-dimensional point p has the capability to query its high-dimensional features F_p from the tri-planes through trilinear interpolation. Subsequently, F_p is directed to a decoder to predict the density σ, color c, and feature f_p of p, facilitating the rendering of the low-resolution image I_l through volume rendering [18]. Finally, both I_l and f_p undergo processing by a super-resolution network, ultimately resulting in the generation of the final high-resolution image I_h.

pix2pix3D [49] currently stands as the state-of-the-art single-modal 3D image synthesis model. It adopts tri-planes as a 3D representation and conducts model training by supervising the difference between the rendered mask and the input mask, as well as the rendered image and the real image. A notable distinction lies in the fact that pix2pix3D only supports a single-modal input, whereas we tackle the more challenging task of multi-modal 3D image synthesis. Our core challenge is to make the model support multi-modal condition input and improve the consistency of the generated results with multi-modal conditions. To address this, we introduce the multi-modal conditional encoders, allowing the model to support multi-modal conditional inputs. We propose an innovative adaptive CLIP loss to improve the consistency between the generated image and the input text. Furthermore, we propose a conditional crossover strategy (CCS) to improve the quality of the generated images.

3.3 Multi-modal 3D generative model

3.3.1 *Multi-modal conditional encoders*

The multi-modal conditional encoders module contains two encoders E_M and E_T, which are responsible for encoding a 2D label map and text respectively. Given a 2D input segmentation map or sketch map $M \in \mathbb{R}^{H\times W\times C}$, we first use a convolutional encoder E_M to map M to the geometry latent code $w_M \in \mathbb{R}^{1\times512}$. Then, a text encoder E_T is used to map the input text and a random latent code $z \sim \mathbb{N}(0, I)$ to the texture W latent code $w_T \in \mathbb{R}^{1\times512}$. We use the text encoder of an adaptive fine-tuned CLIP model as E_T.

$$w_M = E_M(M) \tag{1}$$

$$w_T = E_T(T, z) \tag{2}$$

The first few layers of a GAN network often determine the geometric information of the generated content, while the last few layers determine its texture information [30, 50]. Therefore, we inject w_M into the first seven layers of the generator G and w_T into the last seven layers of the generator G to obtain the generated tri-planes F_{tri}.

3.3.2 Adaptive CLIP fine-tuning

We use CLIP loss to supervise the alignment of generated images with input text. Since the original CLIP model is trained on a large-scale dataset [51], there is a certain gap between this large-scale dataset [51] and the task-specific dataset. Before training Multi3D, we first make adaptive adjustments to the CLIP model to better match our dataset and tasks. We fine-tune the CLIP model on the training set using adversarial loss [51].

$$\mathcal{L}_{\text{finetune}} = \mathcal{L}_{\text{CE}}(E^{I}_{\text{CLIP}}(I) * E^{T}_{\text{CLIP}}(T), \text{labels}) \tag{3}$$

where E^{I}_{CLIP} and E^{T}_{CLIP} are the image encoder and text encoder of the CLIP model respectively. I and T represent real images and text in the training dataset. $*$ represents the matrix multiplication operation. $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss function, labels is an integer array $[0, \cdots, \text{bs} - 1]$, and bs is the training batch size.

3.3.3 Conditional crossover strategy

In order to enhance the expressive ability of the model and improve the reality of the generated results, we propose a *conditional crossover strategy* during training. We set a crossover probability p_c. For a set of training data M and T, we randomly sample a random number p from 0 to 1. If p is greater than the crossover probability p_c, we randomly sample another text T' from the training set to replace the current text T for training. Otherwise, the current training data are not replaced. This improves the image quality for rare samples, such as a bearded man with long hair.

3.3.4 Generative process

Given a 3D spatial point $x \in \mathbb{R}^3$, we can query its features in the tri-planes F_{tri} and obtain the following information through several layers of MLP networks: (i) density value $\sigma \in \mathbb{R}^1$, (ii) color feature $\phi_c \in \mathbb{R}^{32}$, (iii) color value, $c \in \mathbb{R}^3$, (iv) label feature $\phi_m \in \mathbb{R}^{32}$, and (v) label value $m \in \mathbb{R}^K$. If M is a segmentation map, K is the number of categories. If M is a sketch map, K is 1.

$$F_{\text{tri}} = G(w_M, w_T) \tag{4}$$

$$\sigma, \phi_c, c, \phi_m, m = D(F_{\text{tri}}(x)) \tag{5}$$

The low resolution images I'_l and corresponding label maps M'_l are rendered through volume rendering [18]. By sending a ray from the camera pose to each pixel of the imaging plane, we sample N discrete 3D points on the ray. We get I'_l and M'_l using

$$I'_l = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))c_i \tag{6}$$

$$M'_l = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))m_i \tag{7}$$

$$T_i = \exp\Big(- \sum_{j=1}^{i-1} \sigma_i \delta_i \Big) \tag{8}$$

where δ_i is the distance between points the i-th and the $(i+1)$-th.

We adopt two convolutional neural networks S_I, S_M as our super-resolution networks. S_I inputs the low-resolution images I'_l and image features ϕ_c and generates high-resolution images I'_h. S_M inputs the low-resolution labels M'_l and label features ϕ_m and generates high-resolution labels M'_h.

$$I'_h = S_I(I'_l, \phi_c) \tag{9}$$

$$M'_h = S_M(M'_l, \phi_m) \tag{10}$$

3.4 Learning objective

3.4.1 Goals

We design our loss function considering three requirements: alignment, quality, and consistency. Label reconstruction loss and adaptive CLIP loss are used to ensure that the generated three-dimensional object is aligned with the input 2D label map and input text respectively. GAN loss is used to ensure the quality of generated images. Cross-view consistency loss is used to ensure the three-dimensional consistency of the generated object.

3.4.2 Label reconstruction loss

To ensure that the generated image conforms to the geometric constraints of the input label map, we introduce a *label reconstruction loss* $\mathcal{L}_{\text{recon}}$ between the generated label map M'_l, M'_h and the real label map M.

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_l(M, \{M'_l, M'_h\}) \tag{11}$$

where $\mathcal{L}_l$ refers to cross entropy loss when M is a 2D segmentation map, and L1 loss when M is a sketch map.

3.4.3 Adaptive CLIP loss

Adaptive CLIP loss $\mathcal{L}_{\text{CLIP}}$ is adopted to supervise the generation of images that match the input text T.

$$\begin{aligned}\mathcal{L}_{\text{CLIP}} = \arccos^2(E^{I}_{\text{Adap-CLIP}}(I'_l, I'_h),\\ E^{T}_{\text{Adap-CLIP}}(T))\end{aligned} \tag{12}$$

where $E^{I}_{\text{Adap-CLIP}}$ and $E^{T}_{\text{Adap-CLIP}}$ are respectively the image encoder and text encoder of the fine-tuned adaptive CLIP models, which have a smaller gap between images and text for our task.

3.4.4 GAN loss

Adversarial loss is introduced to supervise the training process of Multi3D. We use two discriminators: (i) an image discriminator D_I, trained to ensure the quality of the generated image, and (ii) an image and label alignment discriminator D_{IM}, trained to ensure the alignment and consistency of the generated image and label map. At the same time, we also introduce a gradient penalty to the discriminator to ensure stability of training:

$$\begin{aligned}\mathcal{L}_{D_I} &= \mathbb{E}[1+\exp(D_I(I'_l, I'_h))] \\ &\quad + \mathbb{E}[1+\exp(-D_I(I_l, I_h))] \\ &\quad + \lambda_{I_{\text{reg}}}\mathbb{E}\|\nabla D_I(I_l, I_h)\|^2 \end{aligned} \tag{13}$$

$$\begin{aligned}\mathcal{L}_{D_{IM}} &= \mathbb{E}[1+\exp(D_{IM}(I'_l, I'_h, M'_l, M'_h))] \\ &\quad + \mathbb{E}[1+\exp(-D_{IM}(I_l, I_h, M'_l, M'_h))] \\ &\quad + \lambda_{I_{\text{reg}}}\mathbb{E}\|\nabla D_{IM}(I_l, I_h, M_l, M_h)\|^2 \end{aligned} \tag{14}$$

3.4.5 Cross-view consistency loss

Following pix2pix3D, we introduce cross-view supervision to ensure 3D consistency of the generated objects. In the current input pose p, for the generated tri-planes, we render a label image M^{novel}_l from another novel random pose p'. Then we feed the label map M^{novel}_l into the generator and render the label map M^{proj}_l on the pose p. We use L1 loss to ensure that the projected label map M^{proj}_l remains consistent with M'.

$$L_{\text{cvc}} = \mathcal{L}_1(M'_l, M^{\text{proj}}_l) \tag{15}$$

3.4.6 Overall learning objective

Our overall learning objective is a weighted combination of the above loss functions.

$$\begin{aligned}\mathcal{L} &= \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{cvc}}\mathcal{L}_{\text{cvc}} + \lambda_{\text{CLIP}}\mathcal{L}_{\text{CLIP}} \\ &\quad + \lambda_{D_I}\mathcal{L}_{D_I} + \lambda_{D_{IM}}\mathcal{L}_{D_{IM}} \end{aligned} \tag{16}$$

In our experiments, we empirically set $\lambda_{\text{recon}} = 2$, $\lambda_{\text{cvc}} = 0.0001$, $\lambda_{\text{CLIP}} = 1$, $\lambda_{D_I} = 1$, $\lambda_{D_{IM}} = 0.1$.

4 Experiments

4.1 Datasets

We conducted quantitative and qualitative evaluations on three datasets: CelebAMask-HQ [15] (24,183 training images, 2824 test images), AFHQ-cat [16] (9117 training images, 1013 test images), and shapenet-car [17] (48,3878 training images, 53,764 test images). We use the camera poses and segmentation or sketch labels constructed by pix2pix3D [49] for these three datasets. For the CelebAMask-HQ dataset, we use the text labels from a previous method [24]. Since the AFHQ-cat and shapenet-car datasets do not have text annotations, we used the BLIP model to construct labels for images by asking “what color is this cat/car?”

4.2 Baseline

Since there is no previous method for multi-modal 3D image synthesis, we compare our results to those from the latest single-mode generative model pix2pix3D [49] and modify it to support our new multi-modal tasks. pix2pix3D encodes the segmentation or sketch maps into latent codes that control the geometry and injects them into the first seven layers of the generator, using random noise to inject into the last seven layers of the generator. To support text control, we feed text latent codes encoded by our adaptive CLIP model into the last seven layers of the generator instead of random noise.

4.3 Metrics

Using the test sets of CelebAMask-HQ, AFHQ-cat, and shapenet-car, we generated 10, 30, and 1 sample for each set of inputs by sampling different z to ensure that the total number of generated samples is approximately 30,000 and above. We evaluated performance using three considerations:

- *Quality*: We evaluate Fréchet inception distance (FID) [52], which measures the distance between all the generated images and all the training images.
- *Alignment*: We evaluate alignment using two aspects: (1) *Geometric alignment*. In scenarios where the input is a 2D segmentation image, we compute both the accuracy and mIoU (mean intersection over union) metrics between the generated segmentation image and the input segmentation image. In the cases where the input is a 2D sketch map, we calculate the L1 difference between the generated sketch image and the input sketch; (2) *CLIP score* (*CS*): we use the CLIP model to calculate the cosine similarity between the generated image and the input text, multiplied by 100 to give the CLIP score.

- *Consistency*: We evaluate facial identity consistency (ID) by calculating the mean Arcface [53] cosine similarity of the rendered images of the same generated face from two different random camera poses following EG3D [8].

4.4 Implementation details

When fine-tuning CLIP, we used the AdamW [54] optimizer with a learning rate of 0.0005. When training Multi3D, we used the Adam [55] optimizer with $\beta_1 = 0$ and $\beta_2 = 0.99$ to train the generator and discriminators, with learning rates for G, D_I, D_{IM} of 0.0025, 0.002, and 0.002 respectively. Using eight NVIDIA GeForce 3090 GPUs, we trained the Multi3D model on the CelebAMask-HQ dataset for four days, the AFHQ-cat dataset for three days and four hours, and the ShapeNet-car dataset for three days and ten hours. Fine-tuning the CLIP model introduced negligible overhead. The duration of CLIP training varied across datasets, from half a day to one and a half days.

4.5 Speed

In terms of inference speed, our method achieves nearly real-time framerates at 512^2 resolution. On a single NVIDIA GeForce 3090 GPU, we can render 15 images and corresponding label maps per second.

4.6 Results

We qualitatively demonstrate the generation ability using the CelebAMask-HQ dataset in Fig. 3. Our approach demonstrates the capability to produce images that exhibit a high degree of consistency with the input text, simultaneously ensuring three-dimensional consistency. At the same time, thanks to the conditional crossover strategy, our method can successfully generate challenging results, such as inputting a segmentation image with long hair and inputting the text "This gentleman is in his forties. His face is covered with beard of medium length." Our method generates samples that do not exist in the dataset. At the same time, the results we generate are more diverse. A quantitative comparison is given in Table 1.

To show our generalization ability, we conducted a quantitative comparison with four mainstream

Table 1 Comparison of quality, alignment, and consistency using the CelebAMask-HQ dataset. CS = CLIP score, which evaluates the alignment of generated images and text

Method	Quality	Alignment			Consistency
	FID ↓	CS ↑	acc ↑	mIoU ↑	ID ↑
pix2pix3D	28.53	2.90	0.74	0.40	0.36
Our	**14.72**	**8.52**	**0.80**	**0.45**	**0.53**

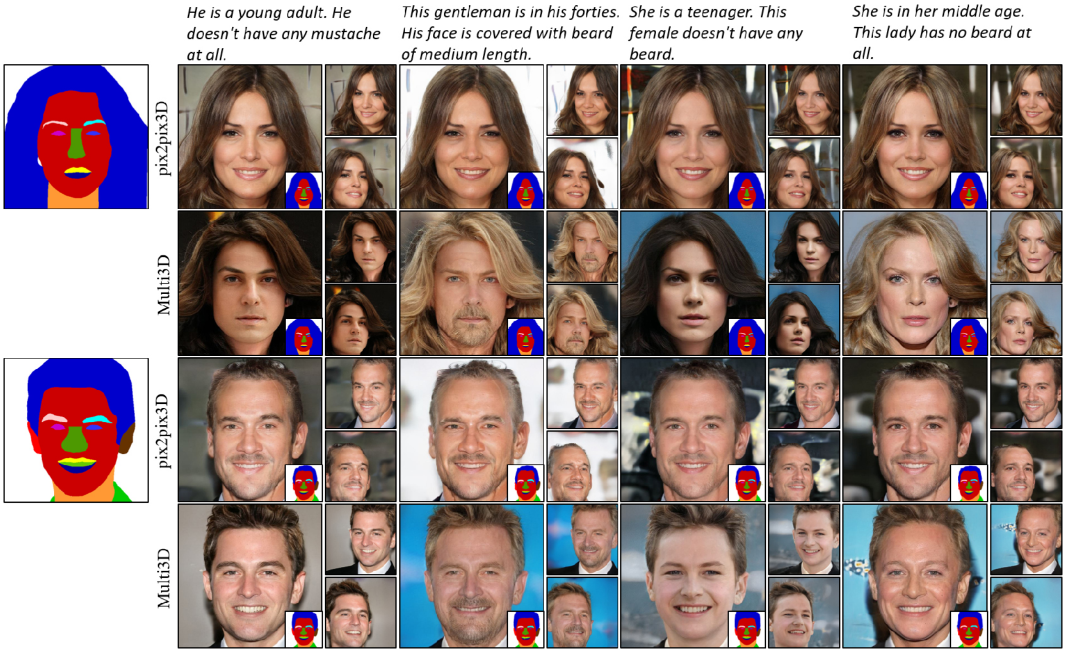
He is a young adult. He doesn't have any mustache at all.
This gentleman is in his forties. His face is covered with beard of medium length.
She is a teenager. This female doesn't have any beard.
She is in her middle age. This lady has no beard at all.
pix2pix3D
Multi3D
pix2pix3D
Multi3D

Fig. 3 3D face generation results driven by both segmentation maps and text. For the same 2D semantic segmentation map input, we demonstrate diverse text-driven face generation results. The results we generate are more consistent with the input segmentation map and text. We can even generate the challenging example of a bearded man with long hair with high quality and 3D consistency.

latest single-modal 3D image generation methods: SEAN [26], SoFGAN [33], Pix2NeRF [48], and pix2pix3D* [49] (the original single-modal version of pix2pix3D). Their FIDs are 32.74, 23.34, 54.23, and 17.11 respectively, while ours is 14.72. This means that even though our task is more challenging, the quality of our generated images is still state-of-the-art. In addition, we also show qualitative results alongside ones from the best single-modal model pix2pix3D* in Fig. 4. It is evident that pix2pix3D* results exhibit inconsistencies with the input mask, exemplified by the generation of incorrect earrings, the absence of an open mouth, and unnatural texture artifacts. Furthermore, the images generated by pix2pix3D* are limited to a single modality. For example, given a mask with long hair, pix2pix3D* can only generate young women. Given a short hair mask, it can only generate young men. In contrast, Multi3D can generate diverse results for different genders, ages, etc.

Figure 5 shows qualitative results when generating 3D cats from segmentation maps and text using the AFHQ-cat dataset. Table 2 quantitatively compares results with those of pix2pxix3D*.

Figure 6 shows the qualitative results of generating 3D cars from sketches and text on the shapenet-car dataset. It can be seen that our method has relatively good decoupling in multiple condition controls, and the shape of the vehicle generated under different text inputs is consistent with the input sketch. However, pix2pix3D has different vehicle shapes for different texts, and even generates wrong geometry. Table 3

Table 2 Quantitative comparison of quality, alignment, and consistency using the AFHQ-cat dataset

Method	Quality	Alignment		
	FID ↓	CS ↑	acc ↑	mIoU ↑
pix2pix3D	28.95	28.08	0.83	0.55
Our	**13.34**	**50.24**	**0.84**	**0.57**

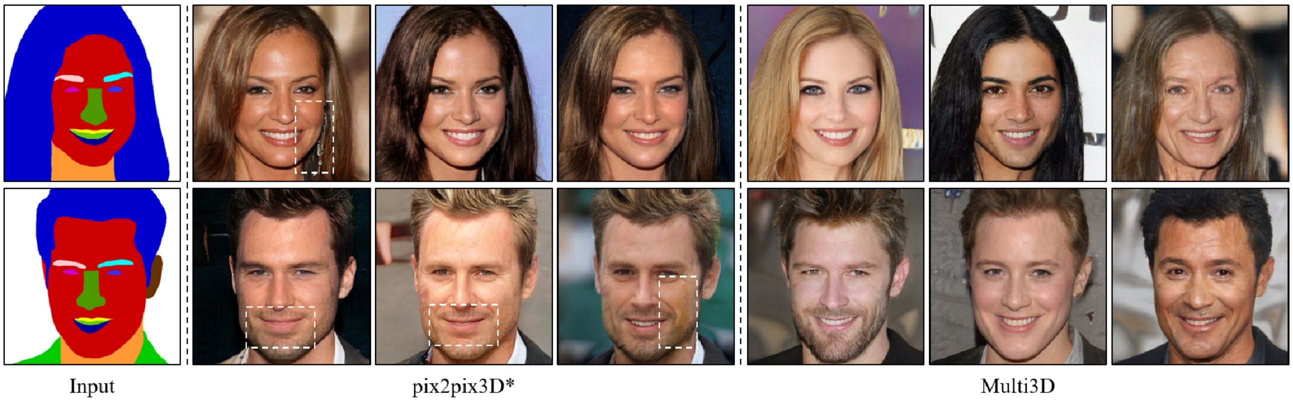

Fig. 4 Results compared to those of pix2pix3D* (the original state-of-the-art single-modal version of pix2pix3D [49]). Its results are not only inconsistent with the input mask, e.g., in terms of incorrect earrings, closed mouths, and unnatural texture artifacts, but are also limited to one mode, only generating young women given long-haired masks and young men given short-haired masks. In contrast, Multi3D provides diverse results across genders and ages.

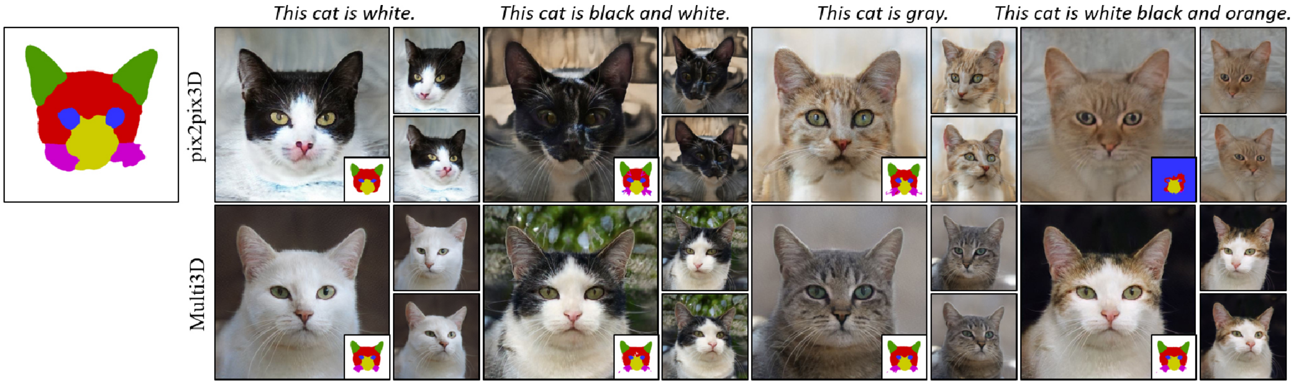
This cat is white.
This cat is black and white.
This cat is gray.
This cat is white black and orange.
pix2pix3D
Multi3D

Fig. 5 3D cat generation results driven by both segmentation maps and text. For the same 2D semantic segmentation map input, we demonstrate diverse text-driven face generation results.

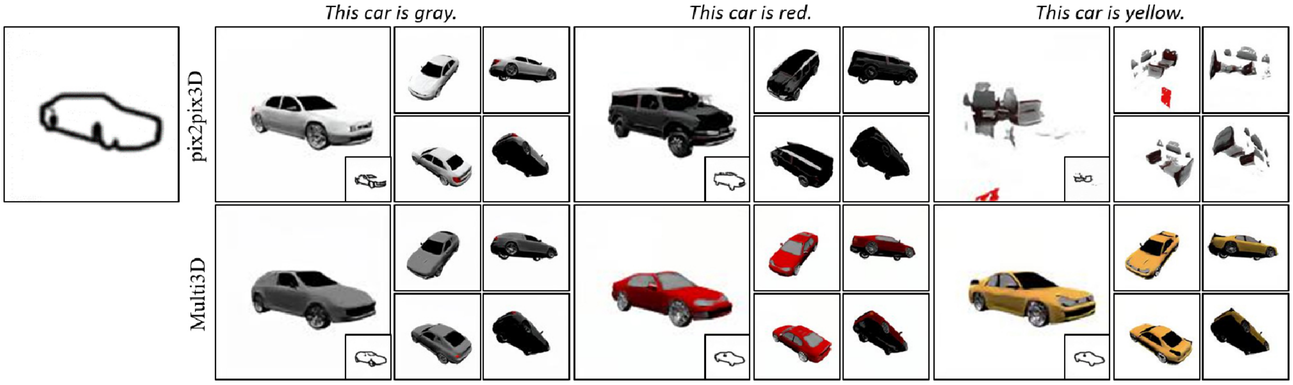

Fig. 6 3D car generation results driven by both sketch maps and text. For the same 2D sketch map input, we demonstrate diverse text-driven car generation results.

Table 3 Quantitative comparison of quality and alignment for the shapenet-car dataset

Method	Quality	Alignment	
	FID ↓	CS ↑	Difference ↓
pix2pix3D	28.37	23.05	**0.16**
Our	**26.74**	**31.71**	**0.16**

gives a quantitative comparison.

4.7 Ablation experiments

We further conducted a full range of ablation experiments, including considering different training strategy settings, crossover probabilities p_c, and types of latent space. All ablation experiments were performed on the CelebAMask-HQ dataset.

4.7.1 Varying training strategy settings

The adaptive CLIP fine-tuning aims to further reduce the gap between images and text on a specific dataset, thereby making the generated image closer to the input text. Example results are shown in Fig. 7: without fine-tuning, the generated results no longer have medium-length beards, which clearly does not conform to the input text. Quantitative metrics are provided in Table 4. The FID and CLIP scores decrease significantly.

Table 4 Ablation experiments with different training strategy settings. Without adaptive CLIP fine-tuning, the generated results no longer have medium-length beards, so clearly do not conform to the input text. Without conditional crossover, the quality of the generated images is lower

Setting	Quality	Alignment			Consistency
	FID ↓	CS ↑	acc ↑	mIoU ↑	ID ↑
w/o ACF	23.26	4.52	0.79	**0.45**	0.49
w/o CCS	17.43	**8.80**	**0.80**	**0.45**	0.51
Full model	**14.72**	8.52	**0.80**	**0.45**	**0.53**

The conditional crossover strategy aims to allow the network to generate more challenging samples, such as bearded men with long hair. As can be seen from Fig. 7, the style-mixing [3, 4, 56] of StyleGAN can make the model generate results consistent with the input text even when there is no conditional crossover strategy. However, as can be seen from Table 4, the FID is lower, as is the quality of the generated images, because the discriminator cannot supervise the samples generated by the cross condition during training.

4.7.2 Conditional crossover probability

We also performed experiments on conditional crossover probabilities p_c. When training a set of conditions M and T, we generated a random number p ranging from 0 to 1. When p was greater than p_c, we randomly selected another set of data M' and T' from the training set, and then we used M and T' for training, otherwise we directly used M and T for training. If p_c is too large, there will be very few crossover samples, which may lead to poor quality of the generated samples. Setting p_c too small will lead to too many crossover samples, which may lead to an excessive amount of data that the model needs to accept. Table 5 shows the effects of p_c taking different values: 0.00, 0.25, 0.50, 0.75. It can be seen that our method obtains the best FID score when $p_c = 0.50$, which we choose in our standard method.

4.7.3 Type of latent space

Our generator network has a total of 14 convolutional layers for injection of style latent codes. We can encode segmentation images, sketches, and text into two alternative latent code spaces through their respective encoders: $W \in \mathbb{R}^{1\times512}$ and $W+ \in \mathbb{R}^{14\times512}$.

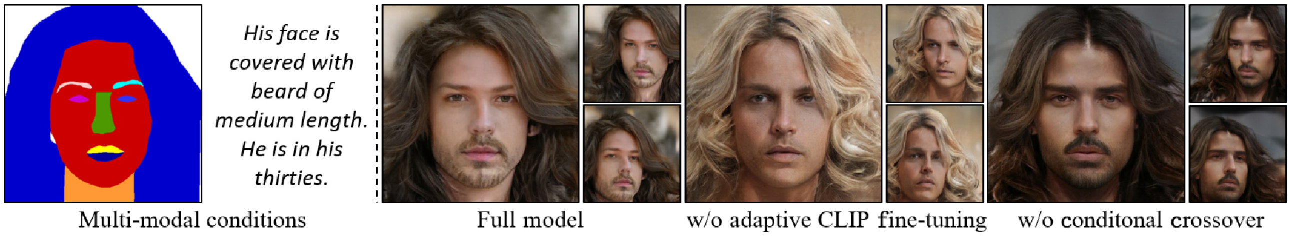
His face is covered with beard of medium length. He is in his thirties.
Multi-modal conditions
Full model
w/o adaptive CLIP fine-tuning
w/o conditonal crossover

Fig. 7 Results of ablation experiments. Without adaptive CLIP fine-tuning, the generated results no longer have medium-length beards, which clearly does not conform to the input text. Without conditional crossover, the texture of the generated results is less realistic, and the generated results look unnatural.

Table 5 Ablation experiments on cross probability p_c. Our method obtains the best FID score when $p_c = 0.50$

p_c	Quality	Alignment			Consistency
	FID ↓	CS ↑	acc ↑	mIoU ↑	ID ↑
0.00	16.35	7.73	0.80	0.44	0.51
0.25	16.21	**8.93**	**0.81**	**0.45**	**0.54**
0.50	**14.72**	8.52	0.80	**0.45**	0.53
0.75	15.12	8.84	0.80	**0.45**	0.52

The W latent code is converted into 14 style latent codes through the repeat operation and then sent to the generator, while the $W+$ latent code is directly sent to the generator. We conducted experiments on both latent code spaces, with results shown in Table 6. W latent code space has better FID and ID, so we choose W latent code space in our standard method.

Table 6 Ablation experiments on the type of latent space

Latent space	Quality	Alignment			Consistency
	FID ↓	CS ↑	acc ↑	mIoU ↑	ID ↑
$W+$	15.80	**8.97**	**0.80**	**0.45**	0.51
W	**14.72**	8.52	**0.80**	**0.45**	**0.53**

5 Conclusions, limitations and future work

We have proposed Multi3D, a multimodal 3D image synthesis model that can generate 3D objects aligned with multiple input conditions. We have proposed a *multimodal condition encoders* module to encode different input conditions. We adopt an *adaptive CLIP fine-tuning* strategy to improve alignment between the generated results and the input text. At the same time, we suggest a *conditional crossover strategy* to improve the quality of generated results. Our method provides state-of-the-art performance on the three datasets: CelebAMask-HQ, AFHQ-cat, and shapenet-car.

One limitation of our approach is that we currently focus on controlling the generation process through two modalities. One mode controls the geometry and the other controls the texture (appearance). Simultaneously supporting three or more modes for hybrid control, including but not limited to segmentation maps, sketches, text, depth maps, attitude maps, etc., is a more challenging research problem. Multimodal 3D-aware image synthesis is an emerging and valuable research direction. We believe our work takes an important step towards multimodal 3D-aware image synthesis. In future, how to complete the task of multi-modal generation for universal three-dimensional objects is one of the most challenging and valuable research directions.

Author contributions

Wenyang Zhou: Methodology, Experiment, Writing—Original Draft. Lu Yuan: Experiment, Writing—Original Draft. Taijiang Mu: Writing—Review and Editing, Supervision.

Acknowledgements

This paper was supported by the National Science and Technology Major Project (Grant No. 2021ZD0112902), the National Natural Science Foundation of China (Project No. 62220106003), a Research Grant from Beijing Higher Institution Engineering Research Center, and Tsinghua–Tencent Joint Laboratory for Internet Innovation Technology.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2672–2680, 2014.

[2] Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In: Proceedings of the International Conference on Learning Representations, 2018.

[3] Karras, T.; Laine, S.; Aila, T. M. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4396–4405, 2019.

[4] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. M. Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8107–8116, 2020.

[5] Isola, P.; Zhu, J. Y.; Zhou, T. H.; Efros, A. A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5967–5976, 2017.

[6] Zhu, J. Y.; Park, T.; Isola, P.; Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2242–2251, 2017.

[7] Park, T.; Liu, M. Y.; Wang, T. C.; Zhu, J. Y. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2332–2341, 2019.

[8] Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B. X.; de Mello, S.; Gallo, O.; Guibas, L.; Tremblay, J.; Khamis, S.; et al. Efficient geometry-aware 3D generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16102–16112, 2022.

[9] OrEl, R.; Luo, X.; Shan, M. Y.; Shechtman, E.; Park, J. J.; Kemelmacher-Shlizerman, I. StyleSDF: High-resolution 3D-consistent image and geometry generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13493–13503, 2022.

[10] Gu, J. T.; Liu, L. J.; Wang, P.; Theobalt, C. StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis. In: Proceedings of the International Conference on Learning Representations, 2022.

[11] Jiang, K. W.; Chen, S. Y.; Liu, F. L.; Fu, H. B.; Gao, L. NeRFFaceEditing: Disentangled face editing in neural radiance fields. In: Proceedings of the SIGGRAPH Asia Conference Papers, Article No. 31, 2022.

[12] Sun, J.; Wang, X.; Shi, Y.; Wang, L.; Wang, J.; Liu, Y. IDE-3D: Interactive disentangled editing for high-resolution 3D-aware portrait synthesis. *ACM Transactions on Graphics* Vol. 41, No. 6, Article No. 270, 2022.

[13] Zhou, W. Y.; Yuan, L.; Chen, S. Y.; Gao, L.; Hu, S. M. LC-NeRF: Local controllable face generation in neural radiance field. *IEEE Transactions on Visualization and Computer Graphics* doi: 10.1109/TVCG.2023.3293653, 2023.

[14] Gao, L.; Liu, F. L.; Chen, S. Y.; Jiang, K. W.; Li, C. P.; Lai, Y. K.; Fu, H. B. SketchFaceNeRF: Sketch-based facial generation and editing in neural radiance fields. *ACM Transactions on Graphics* Vol. 42, No. 4, Article No. 159, 2023.

[15] Lee, C. H.; Liu, Z. W.; Wu, L. Y.; Luo, P. MaskGAN: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5548–5557, 2020.

[16] Choi, Y.; Uh, Y.; Yoo, J.; Ha, J. W. StarGAN v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8185–8194, 2020.

[17] Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. ShapeNet: An information-rich 3D model repository. *arXiv preprint* arXiv:1512.03012, 2015.

[18] Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; Ng, R. NeRF: Representing scenes as neural radiance fields for view synthesis. In: *Computer Vision–ECCV 2020. Lecture Notes in Computer Science, Vol. 12346*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 405–421, 2020.

[19] Müller, T.; Evans, A.; Schied, C.; Keller A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics* Vol. 41, No. 4, Article No. 102, 2022.

[20] Yu, A.; Li, R. L.; Tancik, M.; Li, H.; Ng, R.; Kanazawa, A. PlenOctrees for real-time rendering of neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 5732–5741, 2021.

[21] Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q. H.; Recht, B.; Kanazawa, A. Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5491–5500, 2022.

[22] Shi, Y. C.; Yang, X.; Wan, Y. Y.; Shen, X. H. SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11244–11254, 2022.

[23] Huang, Z. Y.; Peng, Y. C.; Hibino, T.; Zhao, C. Q.; Xie, H. R.; Fukusato, T.; Miyata, K. DualFace: Two-stage drawing guidance for freehand portrait sketching. *Computational Visual Media* Vol. 8, No. 1, 63–77, 2022.

[24] Huang, Z. Q.; Chan, K. C. K.; Jiang, Y. M.; Liu, Z. W. Collaborative diffusion for multi-modal face generation and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6080–6090, 2023.

[25] Liu, X. T.; Wu, W. L.; Li, C. Z.; Li, Y. F.; Wu, H. S. Reference-guided structure-aware deep sketch colorization for cartoons. *Computational Visual Media* Vol. 8, No. 1, 135–148, 2022.

[26] Zhu, P. H.; Abdal, R.; Qin, Y. P.; Wonka, P. SEAN: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5103–5112, 2020.

[27] Xue, Y.; Guo, Y. C.; Zhang, H.; Xu, T.; Zhang, S. H.; Huang, X. L. Deep image synthesis from intuitive user input: A review and perspectives. *Computational Visual Media* Vol. 8, No. 1, 3–31, 2022.

[28] Zhou, W. Y.; Yang, G. W.; Hu, S. M. Jittor-GAN: A fast-training generative adversarial network model zoo based on Jittor. *Computational Visual Media* Vol. 7, No. 1, 153–157, 2021.

[29] Sushko, V.; Schönfeld, E.; Zhang, D.; Gall, J.; Schiele, B.; Khoreva, A. OASIS: Only adversarial supervision for semantic image synthesis. *International Journal of Computer Vision* Vol. 130, No. 12, 2903–2923, 2022.

[30] Xia, W. H.; Yang, Y. J.; Xue, J. H.; Wu, B. Y. TediGAN: Text-guided diverse face image generation and manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2256–2265, 2021.

[31] Wang, T. C.; Liu, M. Y.; Zhu, J. Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8798–8807, 2018.

[32] Patashnik, O.; Wu, Z. Z.; Shechtman, E.; Cohen-Or, D.; Lischinski, D. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2065–2074, 2021.

[33] Chen, A.; Liu, R.; Xie, L.; Chen, Z.; Su, H.; Yu, J. SofGAN: A portrait image generator with dynamic styling. *ACM Transactions on Graphics* Vol. 41, No. 1, Article No. 1, 2022.

[34] Ling, H.; Kreis, K.; Li, D.; Kim, S. W.; Torralba, A.; Fidler, S. EditGAN: High-precision semantic image editing. In: Proceedings of the 35th Conference on Neural Information Processing Systems, 16331–16345, 2021.

[35] Sun, R. Q.; Huang, C.; Zhu, H. L.; Ma, L. Z. Mask-aware photorealistic facial attribute manipulation. *Computational Visual Media* Vol. 7, No. 3, 363–374, 2021.

[36] Wang, C.; Tang, F.; Zhang, Y.; Wu, T. R.; Dong, W. M. Towards harmonized regional style transfer and manipulation for facial images. *Computational Visual Media* Vol. 9, No. 2, 351–366, 2023.

[37] Chen, S. Y.; Su, W. C.; Gao, L.; Xia, S. H.; Fu, H. B. DeepFaceDrawing: Deep generation of face images from sketches. *ACM Transactions on Graphics* Vol. 39, No. 4, Article No. 72, 2020.

[38] Chen, S. Y.; Liu, F. L.; Lai, Y. K.; Rosin, P. L.; Li, C. P.; Gao, L. DeepFaceEditing: Deep face generation and editing with disentangled geometry and appearance control. *ACM Transactions on Graphics* Vol. 40, No. 4, Article No. 90, 2021.

[39] Chan, E. R.; Monteiro, M.; Kellnhofer, P.; Wu, J. J.; Wetzstein, G. 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, 5795–5805, 2021.

[40] Niemeyer, M.; Geiger, A. GIRAFFE: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11448–11459, 2021.

[41] Deng, Y.; Yang, J. L.; Xiang, J. F.; Tong, X. GRAM: Generative radiance manifolds for 3D-aware image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10663–10673, 2022.

[42] Xiang, J. F.; Yang, J. L.; Deng, Y.; Tong, X. GRAM-HD: 3D-consistent image generation at high resolution with generative radiance manifolds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2195–2205, 2023.

[43] Sun, J. X.; Wang, X.; Zhang, Y.; Li, X. Y.; Zhang, Q.; Liu, Y. B.; Wang, J. FENeRF: Face editing in neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7662–7672, 2022.

[44] Chen, Y. D.; Wu, Q. Y.; Zheng, C. X.; Cham, T. J.;

Cai, J. F. Sem2NeRF: Converting single-view semantic masks to neural radiance fields. In: *Computer Vision – ECCV 2022. Lecture Notes in Computer Science, Vol. 13674*. Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; Hassner, T. Eds. Springer Cham, 730–748, 2022.

[45] Jiang, K. W.; Chen, S. Y.; Fu, H. B.; Gao, L. NeRFFaceLighting: Implicit and disentangled face lighting representation leveraging generative prior in neural radiance fields. *ACM Transactions on Graphics* Vol. 42, No. 3, Article No. 35, 2023.

[46] Tang, J. S.; Zhang, B.; Yang, B. X.; Zhang, T.; Chen, D.; Ma, L. Z.; Wen, F. 3DFaceShop: Explicitly controllable 3D-aware portrait generation. *IEEE Transactions on Visualization and Computer Graphics* doi: 10.1109/TVCG.2023.3323578, 2023.

[47] Sun, J. X.; Wang, X.; Wang, L. Z.; Li, X. Y.; Zhang, Y.; Zhang, H. W.; Liu, Y. B. Next3D: Generative neural texture rasterization for 3D-aware head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 20991–21002, 2023.

[48] Cai, S. Q.; Obukhov, A.; Dai, D. X.; Van Gool, L. Pix2NeRF: Unsupervised conditional π-GAN for single image to neural radiance fields translation In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3971–3980, 2022.

[49] Deng, K. L.; Yang, G. S.; Ramanan, D.; Zhu, J. Y. 3D-aware conditional image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4434–445, 2023.

[50] Zhu, P. H.; Abdal, R.; Qin, Y. P.; Wonka, P. Improved StyleGAN embedding: Where are the good latents? *arXiv preprint* arXiv:2012.09036, 2020.

[51] Radford, A.;, Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision In: Proceedings of the 38th International Conference on Machine Learning, 8748–8763, 2021.

[52] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6629–6640, 2017.

[53] Deng, Y.; Yang, J. L.; Xu, S. C.; Chen, D.; Jia, Y. D.; Tong, X. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 285–295, 2019.

[54] Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations, 2019.

[55] Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations, 2015.

[56] Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-free generative adversarial networks. In: Proceedings of the 35th Conference on Neural Information Processing Systems, 852–863, 2021.

Wenyang Zhou is currently a Ph.D. student in the Department of Computer Science and Technology, Tsinghua University. His research interests include computer graphics, 3D-aware generation, and computer vision.

Lu Yuan is currently a master student at Stanford University. Her research interests include computer graphics and computer vision.

Taijiang Mu is currently a research assistant in the Department of Computer Science and Technology, Tsinghua University, where he received his bachelor and doctor degrees in 2011 and 2016, respectively. His research interests include computer graphics, visual media learning, 3D reconstruction, and 3D understanding.