DiffusionGAN3D: Boosting Text-guided 3D Generation and Domain Adaptation by Combining 3D GANs and Diffusion Priors

Biwen Lei, Kai Yu, Mengyang Feng, Miaomiao Cui, Xuansong Xie
Alibaba Group
{biwen.lbw, jinmao.yk, mengyang.fmy, miaomiao.cmm}@alibaba-inc.com,
xingtong.xxs@taobao.com

Figure 1. Some results of the proposed DiffusionGAN3D on different tasks.

Abstract

Text-guided domain adaptation and generation of 3D-aware portraits find many applications in various fields. However, due to the lack of training data and the challenges in handling the high variety of geometry and appearance, the existing methods for these tasks suffer from issues like inflexibility, instability, and low fidelity. In this paper, we propose a novel framework DiffusionGAN3D, which boosts text-guided 3D domain adaptation and generation by combining 3D GANs and diffusion priors. Specifically, we integrate the pre-trained 3D generative models (e.g., EG3D) and text-to-image diffusion models. The former provides a strong foundation for stable and high-quality avatar generation from text. And the diffusion models in turn offer powerful priors and guide the 3D generator finetuning with informative direction to achieve flexible and efficient text-guided domain adaptation. To enhance the diversity in domain adaptation and the generation capability in text-to-avatar, we introduce the relative distance loss and case-specific learnable triplane respectively. Besides, we design a progressive texture refinement module to improve the texture quality for both tasks above. Extensive experiments demonstrate that the proposed framework achieves excellent results in both domain adaptation and text-to-avatar tasks, outperforming existing methods in terms of generation quality and efficiency. The project homepage is at https://younglbw.github.io/DiffusionGAN3D-homepage/.

1. Introduction

3D portrait generation and stylization find a vast range of applications in many scenarios, such as games, advertisements, and film production. While extensive works [4, 7, 9, 17] yield impressive results on realistic portrait generation, the performance on generating stylized, artistic, and text-guided 3D avatars is still unsatisfying due to the lack of 3D training data and the difficulties in modeling highly variable geometry and texture.

Some works [2, 25, 26, 47, 51, 53, 56] perform transfer learning on a pre-trained 3D GAN generator to achieve 3D stylization, which relies on a large number of stylized images and strictly aligned camera poses for training. [2, 47] leverage existing 2D-GAN trained on a specific domain to

synthesize training data and implement finetuning with adversarial loss. In contrast, [25, 26, 51] utilize text-to-image diffusion models to generate training datasets in the target domain. This enables more flexible style transferring but also brings problems like pose bias, tedious data processing, and heavy computation costs. Unlike these adversarial finetuning based methods, StyleGAN-Fusion [48] adopts SDS [37] loss as guidance of text-guided adaptation of 2D and 3D generators, which gives a simple yet effective way to fulfill domain adaptation. However, it also suffers from limited diversity and suboptimal text-image correspondence.

The recently proposed Score Distillation Sampling (SDS) algorithm [37] exhibits impressive performance in text-guided 3D generation. Introducing diffusion priors into the texture and geometry modeling notably reduces the training cost and offers powerful 3D generation ability. However, it also leads to issues like unrealistic appearance and Janus (multi-face) problems. Following [37], massive works [5, 21, 27, 30, 49, 50, 52] have been proposed to enhance the generation quality and stability. Nevertheless, the robustness and visual quality of the generated model are still far less than the current generated 2D images.

Based on the observations above, we propose a novel two-stage framework DiffusionGAN3D to boost the performance of 3D domain adaptation and text-to-avatar tasks by combining 3D generative models and diffusion priors, as shown in Fig. 2. For the text-guided 3D **Domain Adaptation** task, we first leverage diffusion models and adopt SDS loss to finetune a pre-trained EG3D-based model [4, 7, 9] with random noise input and camera views. The relative distance loss is introduced to deal with the loss of diversity caused by the SDS technique. Additionally, we design a diffusion-guided reconstruction loss to adapt the framework to local editing scenarios. Then, we extend the framework to **Text-to-Avatar** task by finetuning 3D GANs with a fixed latent code that is obtained guided by CLIP [38] model. During optimization, a case-specific learnable triplane is introduced to strengthen the generation capability of the network. To sum up, in our framework, the diffusion models offer powerful text-image priors, which guide the domain adaptation of the 3D generator with informative direction in a flexible and efficient way. In turn, 3D GANs provide a strong foundation for text-to-avatar, enabling stable and high-quality avatar generation. Last but not least, taking advantage of the powerful 2D synthesis capability of diffusion models, we propose a **Progressive Texture Refinement** module as the second stage for these two tasks above, which significantly enhances the texture quality. Extensive experiments demonstrate that our method exhibits excellent performance in terms of generation quality and stability on 3D domain adaptation and text-to-avtar tasks, as shown in Fig. 1.

Our main contributions are as follows:

(A) We achieve text-guided 3D domain adaptation in high quality and diversity by combining 3D GANs and diffusion priors with the assistance of the relative distance loss.
(B) We adapt the framework to a local editing scenario by designing a diffusion-guided reconstruction loss.
(C) We achieve high-quality text-to-avatar in superior performance and stability by introducing the case-specific learnable triplane.
(D) We propose a novel progressive texture refinement stage, which fully exploits the image generation capabilities of the diffusion models and greatly enhances the quality of texture generated above.

2. Related Work

Domain Adaptation of 3D GANs. The advancements in 3D generative models [4, 6, 7, 9, 11, 13, 14, 29, 35, 45] have enabled geometry-aware and pose-controlled image generation. Especially, EG3D [7] utilizes triplane as 3D representation and integrates StyleGAN2 [24] generator with neural rendering [33] to achieve high-quality 3D shapes and view-consistency image synthesis, which facilitates the downstream applications such as 3D stylization, GAN inversion [28]. Several works [2, 22, 53, 56] achieve 3D domain adaptation by utilizing stylized 2D generator to synthesize training images or distilling knowledge from it. In contrast, [25, 26, 51] leverage the powerful diffusion models to generate training datasets in the target domain and accomplish text-guided 3D domain adaptation with great performance. Though achieving impressive results, these adversarial learning based methods above suffer from issues such as pose bias, tedious data processing, and heavy computation cost. Recently, non-adversarial finetuining methods [3, 12, 48] also exhibit great promise in text-guided domain adaptation. Especially, StyleGAN-Fusion [48] adopts SDS loss as guidance for the adaptation of 2D generators and 3D generators. It achieves efficient and flexible text-guided domain adaptation but also faces the problems of limited diversity and suboptimal text-image correspondence.

Text-to-3D Generation. In recent years, text-guided 2D image synthesis [10, 41–43, 55] achieve significant progress and provide a foundation for 3D generation. Prior works, including CLIP-forge [44], CLIP-Mesh [34], and DreamFields [20], employ CLIP [38] as guidance to optimize 3D representations such as meshes and NeRF [33]. DreamFusion [37] first proposes score distillation sampling (SDS) loss to utilize a pre-trained text-to-image diffusion model to guide the training of NeRF. It is a pioneering work and exhibits great promise in text-to-3d generation, but also suffers from over-saturation, over-smoothing, and Janus (multi-face) problem. Subsequently, extensive improvements [30, 39, 49, 50] over DreamFusion have been introduced to address these issues. ProlificDreamer [50] proposes variational score distillation (VSD) and produces

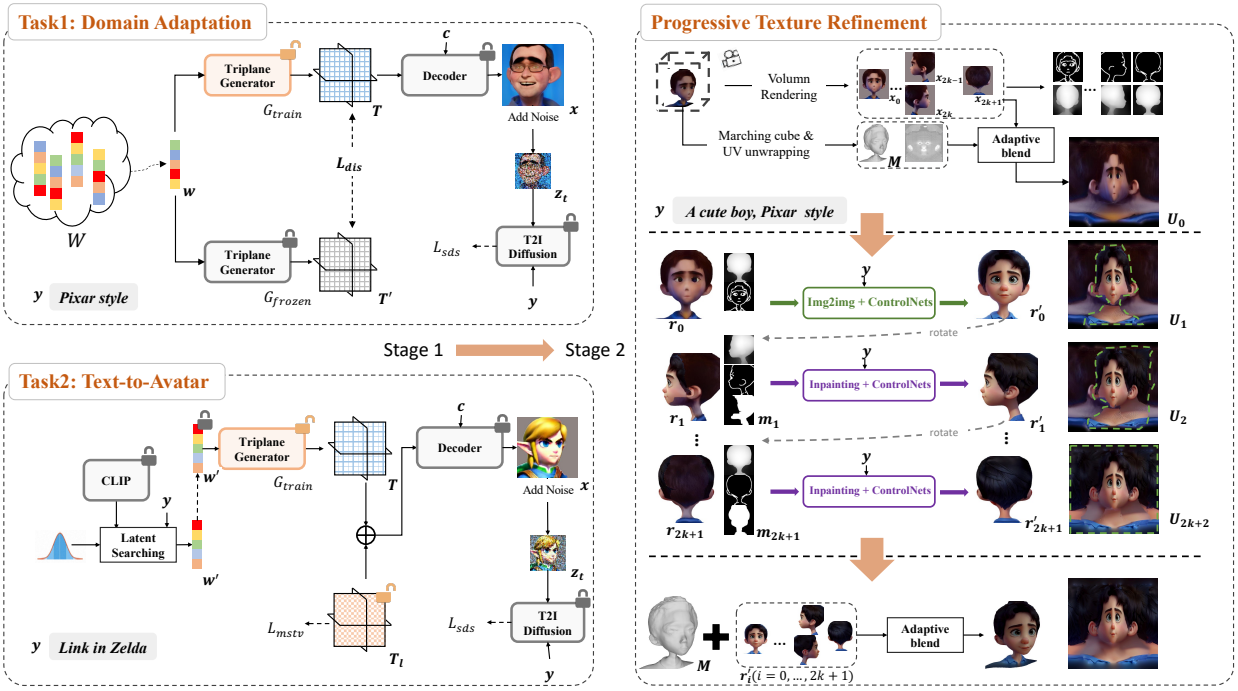


Figure 2. Overview of the proposed two-stage framework DiffusionGAN3D.

high-fidelity texture results. Magic3D [30] adopts a coarse-to-fine strategy and utilizes DMTET [46] as the 3D representation to implement texture refinement through SDS loss. Despite yielding impressive progress, the appearance of their results is still unsatisfying, existing issues such as noise [50], lack of details [30, 49], multi-view inconsistency [8, 40]. Moreover, these methods still face the problem of insufficient robustness and incorrect geometry. When it comes to avatar generation, these shortcomings can be more obvious and unacceptable.

Text-to-Avatar Generation. To handle 3D avatar generation from text, extensive approaches [5, 18, 19, 21, 27, 54] have been proposed. Avatar-CLIP [18] sets the foundation by initializing human geometry with a shape VAE and employing CLIP to guide geometry and texture modeling. DreamAvatar [5] and AvatarCraft [21] fulfill robust 3D avatar creation by integrating the human parametric model SMPL [31] with pre-trained text-to-image diffusion models. DreamHuman [27] further introduces a camera zoom-in strategy to refine the local details of 6 important body regions. Recently, AvatarVerse [52] and a concurrent work [36] employ DensePose-conditioned ControlNet [55] for SDS guidance to realize more stable avatar creation and pose control. Although these methods exhibit quite decent results, weak SDS guidance still hampers their performance in multi-view consistency and texture fidelity.

3. Methods

In this section, we present DiffusionGAN3D, which boosts the performance of 3D domain adaptation and text-to-avatar by combining and taking advantage of 3D GANs and diffusion priors. Fig. 2 illustrates the overview of our framework. After introducing some preliminaries (Sec. 3.1), we first elaborate our designs in diffusion-guided 3D domain adaptation (Sec. 3.2) , where we propose a relative distance loss to resolve the problem of diversity loss caused by SDS. Then we extend this architecture and introduce a case-specific learnable triplane to fulfill 3D-GAN based text-to-avatar (Sec. 3.3). Finally, we design a novel progressive texture refinement stage (Sec. 3.4) to improve the detail and authenticity of the texture generated above.

3.1. Preliminaries

EG3D [7] is a SOTA 3D generative model, which employ triplane as 3D representation and integrate StyleGAN2 [24] generator with neural rendering [33] to achieve high quality 3D shapes and pose-controlled image synthesis. It is composed of (1) a mapping network that projects the input noise to the latent space W, (2) a triplane generator that synthesizes the triplane with the latent code as input, and (3) a decoder that includes a triplane decoder, volume rendering module and super-resolution module in sequence. Given a triplane and camera poses as input, the decoder generates high-resolution images with view consistency.

Score Distillation Sampling (SDS), proposed by DreamFusion [7], utilizes a pre-trained diffusion model ϵ_ϕ as prior for optimization of a 3D representation θ. Given an image $\boldsymbol{x} = g(\theta)$ that is rendered from a differentiable model g, we add random noise $\boldsymbol{\epsilon}$ on $\boldsymbol{x}$ at noise level t to obtain a noisy image $\boldsymbol{z_t}$. The SDS loss then optimizes θ by minimizing

the difference between the predicted noise $\epsilon_\phi(\boldsymbol{z_t}; \boldsymbol{y}, t)$ and the added noise $\boldsymbol{\epsilon}$, which can be presented as:

$$\nabla_\theta L_{SDS}(\phi, g_\theta) = \mathbb{E}_{t,\epsilon}\left[w_t\left(\boldsymbol{\epsilon}_\phi(\boldsymbol{z_t}; \boldsymbol{y}, t) - \boldsymbol{\epsilon}\right)\frac{\partial \boldsymbol{x}}{\partial \theta}\right], \quad (1)$$

where $\boldsymbol{y}$ indicates the text prompt and w_t denotes a weighting function that depends on the noise level t.

3.2. Diffusion-Guided 3D Domain Adaptation

Due to the difficulties in obtaining high-quality pose-aware data and model training, adversarial learning methods for 3D domain adaptation mostly suffer from the issues of tedious data processing and mode collapse. To address that, we leverage diffusion models and adopt the SDS loss to implement transfer learning on an EG3D-based 3D GAN to achieve efficient 3D domain adaptation, as shown in Fig. 2.

Given a style code $\boldsymbol{w}$ generated from noise $\boldsymbol{z} \sim N(0, 1)$ through the fixed mapping network, we can obtain the triplane $\boldsymbol{T}$ and the image $\boldsymbol{x}$ rendered in a view controlled by the input camera parameters $\boldsymbol{c}$ using the triplane generator and decoder in sequence. Then SDS loss (Sec. 3.1) is applied on $\boldsymbol{x}$ to finetune the network. Different from DreamFusion which optimizes a NeRF network to implement single object generation, we shift the 3D generator with random noise and camera pose to achieve domain adaptation guided by text $\boldsymbol{y}$. During optimization, all parameters of the framework are frozen except the triplane generator. We find that the gradient provided by SDS loss is unstable and can be harmful to some other well-trained modules such as the super-resolution module. Besides, freezing the mapping network ensures that the latent code $\boldsymbol{w}$ lies in the same domain during training, which is a crucial feature that can be utilized in the diversity preserving of the 3D generator.

Relative Distance Loss. The SDS loss provides diffusion priors and achieves text-guided domain adaptation of 3D GAN in an efficient way. However, it also brings the problem of diversity loss as illustrated in [48]. To deal with that, [48] proposes the directional regularizer to regularize the generator optimization process, which improves the diversity to a certain extent. However, it also limits the domain shifting, facing a trade-off between diversity and the degree of style transfer. To address this, we propose a relative distance loss. As shown in Fig. 3, considering two style codes $\boldsymbol{w_i}$ and $\boldsymbol{w_j}$ which are mapping from two different noise $\boldsymbol{z_i}$ and $\boldsymbol{z_j}$, we project them into the original triplane domain ($\boldsymbol{T_i'}$, $\boldsymbol{T_j'}$) and the finetuned one ($\boldsymbol{T_i}$, $\boldsymbol{T_j}$) using a frozen triplane generator G_{frozen} and the finetuned triplane generator G_{train}, respectively. Note that, since the mapping network is frozen during training in our framework, $\boldsymbol{T_i}$ and $\boldsymbol{T_i'}$ (same for $\boldsymbol{T_j}$ and $\boldsymbol{T_j'}$) share the same latent code and ought to be close in context. Thus, we model the relative distance of these two samples in triplane space and formulate the relative distance loss L_{dis} as:

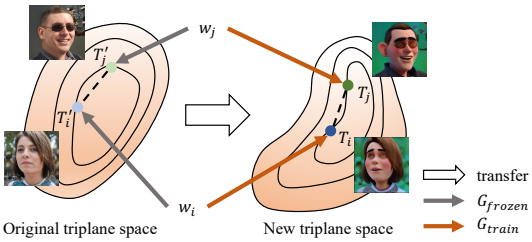


Figure 3. An illustration of the relative distance loss.

$$L_{dis} = abs(\frac{||\boldsymbol{T_i'} - \boldsymbol{T_j'}||^2}{||\boldsymbol{T_i} - \boldsymbol{T_j}||^2} - 1). \quad (2)$$

In this function, guided by the original network, the samples in the triplane space are forced to maintain distance from each other. This prevents the generator from collapsing to a fixed output pattern. Note that it only regularizes the relative distance between different samples while performing no limitation to the transfer of the triplane domain itself. Extensive experiments in Sec. 4 demonstrate that the proposed relative distance loss effectively improves the generation diversity without impairing the degree of stylization.

Diffusion-guided Reconstruction Loss. Despite the combination of SDS loss and the proposed relative distance loss is adequate for most domain adaptation tasks, it still fails to handle the local editing scenarios. A naive solution is to perform reconstruction loss between the rendered image and the one from the frozen network. However, it will also inhibit translation of the target region. Accordingly, we propose a diffusion-guided reconstruction loss especially for local editing, which aims to preserve non-target regions while performing 3D editing on the target region. We found that the gradient of SDS loss has a certain correlation with the target area, especially when the noise level t is large, as shown in Fig. 4. To this end, we design a diffusion-guided reconstruction loss L_{diff} that can be presented as:

$$\gamma = abs(w_t(\boldsymbol{\epsilon}_\phi(\boldsymbol{z_t}; \boldsymbol{y}, t) - \boldsymbol{\epsilon})), \quad (3)$$

$$L_{diff} = t||(\boldsymbol{x} - \boldsymbol{x'}) \odot \left[\boldsymbol{J} - h(\frac{\boldsymbol{\gamma}}{max(\boldsymbol{\gamma})})\right]||^2, \quad (4)$$

where $\boldsymbol{\gamma}$ is the absolute value of the gradient item in Eq. 1, h represents the averaging operation in the feature dimension, $\boldsymbol{J}$ is the matrix of ones having the same spatial dimensions as the output of h, $\boldsymbol{x'}$ denotes the output image of the frozen network under the same noise and camera parameters $\boldsymbol{x}$, $\odot$ indicates the Hadamard product. The latter item of the $\odot$ operation can be regarded as an adaptive mask indicating the non-target region. Compared with ordinary reconstruction loss, the proposed diffusion-guided reconstruction loss alleviates the transfer limitation of the target region. Although the gradient of SDS loss in a single iteration contains a lot of noise and is inadequate to serve as an accurate mask,

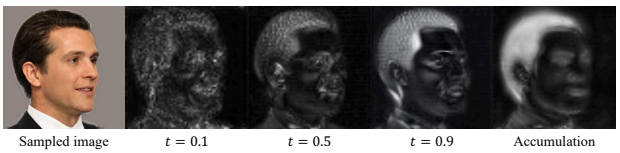


Figure 4. Visualizations of the gradient response of SDS loss at different noise levels, given the text "a man with green hair".

it can also provide effective guidance for network learning with the accumulation of iterations as shown in Fig. 4. The ablation experiment in Sec. 4 also proves its effectiveness.

To sum up, we can form the loss functions for normal domain adaptation and local editing scenario as $L_{adaptation} = L_{sds} + \lambda_1 L_{dis}$ and $L_{editing} = L_{sds} + \lambda_2 L_{diff}$, respectively, where λ_1 and λ_2 are the weighting coefficients.

3.3. 3D-GAN Based Text-to-Avatar

Due to the lack of 3D priors , most text-to-3D methods cannot perform stable generation, suffering from issues such as Janua (multi-face) problem. To this end, we extend the framework proposed above and utilize the pre-trained 3D GAN as a strong base generator to achieve robust text-guided 3D avatar generation. As shown in Fig. 2, we first implement latent searching to obtain the latent code that is contextually (gender, appearance, etc.) close to the text input. Specifically, we sample k noise $\boldsymbol{z_1}, ..., \boldsymbol{z_k}$ and select one single noise $\boldsymbol{z_i}$ that best fits the text description according to the CLIP loss between the corresponding images synthesized by the 3D GAN and the prompt. The CLIP loss is further used to finetune the mapping network individually to obtain the optimized latent code $\boldsymbol{w'}$ from $\boldsymbol{z_i}$. Then, $\boldsymbol{w'}$ is fixed during the following optimization process.

Case-specific learnable triplane. One main challenge of the text-to-avatar task is how to model the highly variable geometry and texture. Introducing 3D GANs as the base generator provides strong priors and greatly improves stability. However, it also loses the flexibility of the simple NeRF network, showing limited generation capability. Accordingly, we introduce a case-specific learnable triplane $\boldsymbol{T_l}$ to enlarge the capacity of the network, as shown in Fig. 2. Initialized with the value of 0, $\boldsymbol{T_l}$ is directly added to $\boldsymbol{T}$ as the input of subsequent modules. Thus, the trainable part of the network now includes the triplane generator G_{train} and $\boldsymbol{T_l}$. The former achieves stable transformation, while the latter provides a more flexible 3D representation. Due to the high degree of freedom of $\boldsymbol{T_l}$ and the instability of SDS loss, optimizing $\boldsymbol{T_l}$ with SDS loss alone will bring a lot of noise, resulting in unsmooth results. To this end, we adopt the total variation loss [23] and expand it to a multi-scale manner L_{mstv} to regularize $\boldsymbol{T_l}$ and facilitate more smoothing results. In general, the loss function for text-to-avatar task can be presented as: $L_{avatar} = L_{sds} + \lambda_3 L_{mstv}$.

Note that, the proposed framework is only suitable for the generation of specific categories depending on the pre-trained 3D GAN, such as head (PanoHead [4]) and human body (AG3D [9]). Nevertheless, extensive experiments show that our framework can well adapt to avatar generation with large domain gaps, benefiting from the strong 3D generator and the case-specific learnable triplane.

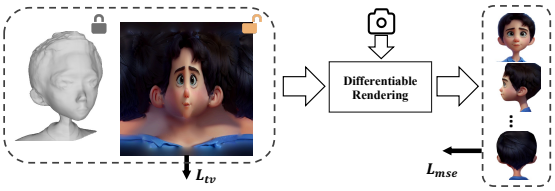


Figure 5. The details of the proposed adaptive blend module.

3.4. Progressive Texture Refinement

The SDS exhibits great promise in geometry modeling but also suffers from texture-related problems such as over-saturation and over-smoothing. How can we leverage the powerful 2D generation ability of diffusion models to improve the 3D textures? In this section, we propose a progressive texture refinement stage, which significantly enhances the texture quality of the results above through explicit texture modeling, as shown in Fig. 2.

Adaptive Blend Module. Given the implicit fields obtained from the first stage, we first implement volume rendering under uniformly selected $2k + 2$ azimuths and j elevations (we set the following j to 1 for simplicity) to obtain multi-view images $\boldsymbol{x_i}, ..., \boldsymbol{x_{2k+1}}$. Then the canny maps and depth maps of these images are extracted for the following image translation. Meanwhile, we perform marching cube [32] and the UV unwrapping [1] algorithm to obtain the explicit mesh $\boldsymbol{M}$ and corresponding UV coordinates (in head generation, we utilize cylinder unwrapping for better visualization). Furthermore, we design an adaptive blend module to project the multi-view renderings back into a texture map through differentiable rendering. Specifically, as shown in Fig. 5, the multi-view reconstruction loss L_{mse} and total variation loss L_{tv} are adopted to optimize the texture map that is initialized with zeros. Compared to directly implementing back-projection, the proposed adaptive blending module produces smoother and more natural textures in spliced areas of different images without compromising texture quality. This optimized UV texture $\boldsymbol{U_0}$ serves as an initialization for the following texture refinement stage.

Progressive Refinement. Since we have already obtained the explicit mesh and the multi-view renderings, a natural idea is to perform image-to-image on the multi-view renderings using diffusion models to optimize the texture. However, it neglects that the diffusion model cannot guarantee the consistency of image translation between different views, which may result in discontinuous texture. To this end, we introduce a progressive inpainting strategy to address this issue. Firstly, we employ a pre-trained text-

to-image diffusion model and ControlNets [55] to implement image-to-image translation guided by the prompt y on the front-view image r_0 that is rendered from M and U_0. The canny and depth extracted above are introduced to ensure the alignment between r_0 and the resulting image r'_0. Then we can obtain the partially refined texture map U_1 by projecting r'_0 into U_0. Next, we rotate the mesh coupled with T_1 (or change the camera view) and render a new image r_1, which is refined again with the diffusion model to get r'_1 and U_2. Differently, instead of image-to-image, we apply inpainting on r_1 with mask m_1 in this translation, which maintains the refined region and thus improves the texture consistency between the adjacent views. Note that the masks $m_1, ..., m_{2k+1}$ indicate the unrefined regions and are dilated to facilitate smoother results in inpainting. Through progressively performing rotation and inpainting, we manage to obtain consistent multi-view images $r'_0, ..., r'_{2k+1}$ that are refined by the diffusion model. Finally, we apply the adaptive blend module again on the refined images to yield the final texture. By implementing refinement on the explicit texture, the proposed stage significantly improves the texture quality in an efficient way.

4. Experiments

4.1. Implementation Details

Our framework is built on an EG3D-based model in the first stage. Specifically, we implement 3D domain adaptation on PanoHead, EG3D-FFHQ, and EG3D-AFHQ for head, face, and cat, respectively. For text-to-avatar tasks, PanoHead and AG3D are adopted as the base generators for head and body generation. We employ StableDiffusion v2.1 as our pre-trained text-to-image model. In the texture refinement stage, StableDiffusion v1.5 coupled with ControlNets are utilized to implement image-to-image and inpainting. More details about the parameters and training setting are specified in supplementary materials.

4.2. Qualitative Comparison

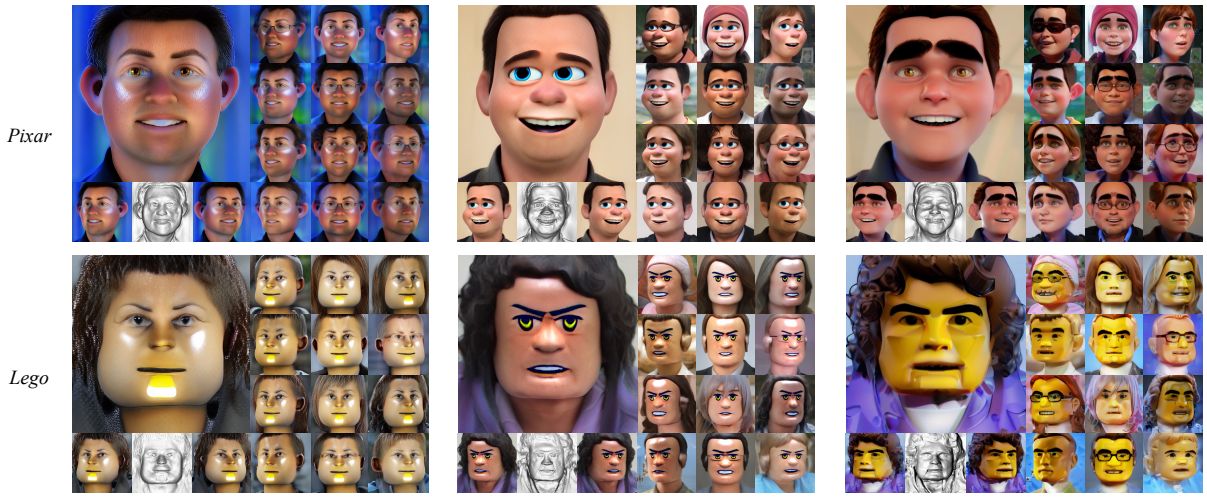
For 3D Domain adaptation, we evaluate our model with two powerful baselines: StyleGAN-NADA* [12] and StyleGAN-Fusion [48] for text-guided domain adaptation of 3D GANs, where * indicates the extension of the method to 3D models. For a fair comparison, we use the same prompts as guidance for all the methods. Besides, the visualization results of different methods are sampled from the same random noise. As shown in Fig. 6, the naive extension of StyleGAN-NADA* for EG3D exhibits poor results in terms of diversity and image quality. StyleGAN-Fusion achieves decent 3D domain adaptation and exhibits a certain diversity. However, the proposed regularizer of StyleGAN-Fusion also hinders itself from large-gap domain transfer, resulting in a trade-off between the degree of stylization and diversity. As Fig. 6 shows that the generated faces of StyleGAN-Fusion lack diversity and details, and the hair and clothes suffer from inadequate stylization. In contrast, our method exhibits superior performance in diversity, image quality, and text-image correspondence.

For text-to-avatar task, We present qualitative comparisons with several general text-to-3D methods (DreamFusion [37], ProlificDreamer [50], Magic-3D [30]) and avatar generation methods (DreamAvatar [5], DreamHuman [27], AvatarVerse [52]). The former three methods are implemented using the official code and the results of the rest methods are obtained directly from their project pages. As shown in Fig. 7, DreamFusion shows inferior performance in avatar generation, suffering from over-saturation, Janus (multi-face) problem, and incorrect body parts. ProlificDreamer and Magic-3D improve the texture fidelity to some extent but still face the problem of inaccurate and unsmooth geometry. Taking advantage of the human priors obtained from the SMPL model or DensePose, these text-to-avatar methods achieve stable and high-quality avatar generation. However, due to that the SDS loss requires a high CFG [16] value during optimization, the texture fidelity and authenticity of their results are still unsatisfying. In comparison, the proposed method achieves stable and high-fidelity avatar generation simultaneously, making full use of the 3D GANs and diffusion priors. Please refer to the supplementary materials for more comparisons.

4.3. Quantitative Comparison

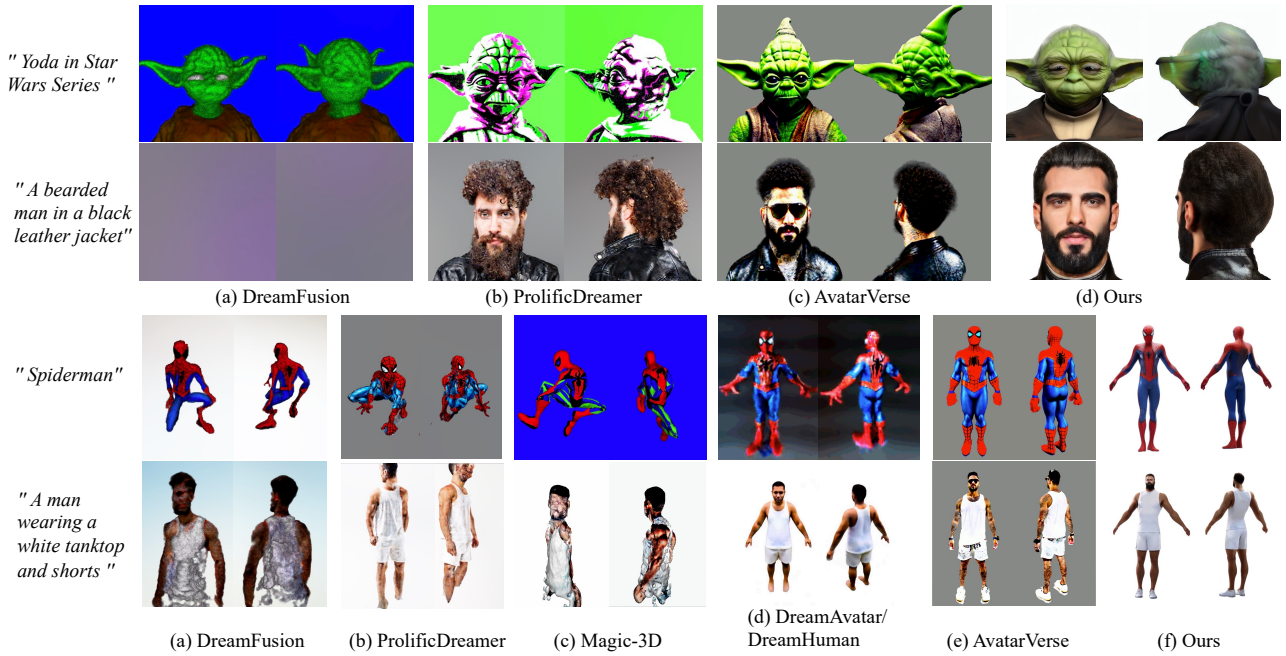
We quantitatively evaluate the above baselines and our method on 3D domain adaptation through FID [15] comparison and user study. Specifically, all methods are employed to conduct domain adaptation on EG3D-face and EG3D-cat with both four text prompts, respectively. For each text prompt, we utilize the text-to-image diffusion model to generate 2000 images with different random seeds as the ground truth for FID calculation. In the user study, 12 volunteers were invited to rate each finetuned model from 1 to 5 based on three dimensions: text-image correspondence, image quality, and diversity. As shown in Table 1, the proposed methods achieve lower FID scores than other baselines, which indicates superior image fidelity. Meanwhile, the user study demonstrates that our method outperforms the other two methods, especially in terms of image quality and diversity.

For text-to-avatar, we also conducted a user study for quantitative comparison. Since AvatarVerse and DreamAvatar have not released their code yet, while DreamHuman provided extensive results on the project page. So we compare our method with DreamHuman for full-body generation. Besides, DreamFusion, ProlificDreamer, and Magic3D are involved in the comparison of head (10 prompts) and full-body (10 prompts) generation both. We



(a) StyleGAN-NADA* (b) StyleGAN-Fusion (c) Ours

Figure 6. The qualitative comparisons on 3D domain adaptation (applied on EG3D-FFHQ [7]).



(a) DreamFusion (b) ProlificDreamer (c) AvatarVerse (d) Ours

(a) DreamFusion (b) ProlificDreamer (c) Magic-3D (d) DreamAvatar/ DreamHuman (e) AvatarVerse (f) Ours

Figure 7. Visual comparisons on text-to-avatar task. The first two rows are the results of 'head' and the rest are the results of 'body'.

request the 12 volunteers to vote for their favorite results based on texture and geometry quality, where all the results are presented as rendered rotating videos. The final rates presented in Table 2 show that the proposed method performs favorably against the other approaches.

Table 1. Quantitative comparison on 3D domain adaptation task.

Methods	Metric	User Study		
	FID ↓	text-corr ↑	quality ↑	diversity ↑
StyleGAN-NADA*	136.2	2.619	2.257	1.756
StyleGAN-Fusion	53.6	3.465	3.255	2.978
Ours	**28.5**	**3.725**	**3.758**	**3.416**

Table 2. User preference on text-to-avatar generation.

	DreamFusion	ProlificDreamer	Magic3D	DreamHuman	Ours
head	1.1%	11.7%	6.7%	N.A.	**80.5%**
body	0.6%	8.3%	5.6%	18.9%	**66.6%**

4.4. Ablation Study

On progressive texture refinement. Since we utilize cylinder unwrapping for head texture refinement, a naive idea is to conduct image-to-image on the UV texture directly to refine it. However, the result in Fig. 8 (b) shows that this method tends to yield misaligned texture, let alone be applied to fragmented texture maps. We also attempt to

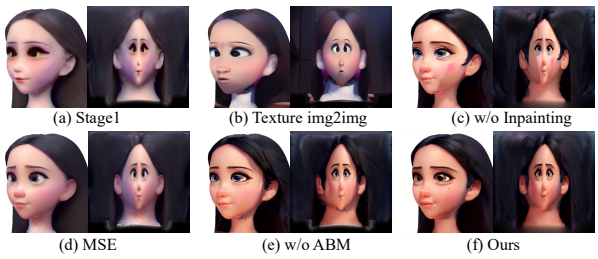


Figure 8. Ablation study of the texture refinement.

Figure 9. Ablation study of the relative distance loss.

replace all the inpainting operations with image-to-image translation, and the results in Fig. 8 (c) show that this will cause the discontinuity problem. The refining strategy proposed in [49] is also compared, where texture is progressively optimized using MSE loss between the randomly rendered images and the corresponding image-to-image results. The results in Fig. 8 (d) show that it fails to generate high-frequency details. The comparison between (e) and (f) in Fig. 8 proves the effectiveness of the proposed adaptive blend module (ABM) in smoothing the texture splicing region. By contrast, the proposed progressive texture refinement strategy significantly improves the texture quality.

On relative distance loss. As shown in Fig. 9, if adopting SDS loss alone for domain adaptation, the generator will tend to collapse to a fixed output pattern, losing its original diversity. In contrast, the proposed relative distance loss effectively preserves the diversity of the generator without sacrificing the stylization degree.

On diffusion-guided reconstruction loss. The results in Fig 10 show that the SDS loss tends to perform global transfer. Regular reconstruction loss helps maintain the whole structure, but also stem the translation of the target area. By contrast, the model trained with our diffusion-guided reconstruction loss achieves proper editing.

On additional learnable triplane. To prove the necessity of the proposed case-specific learnable triplane, we finetune the network with SDS loss without adding it, given a challenging prompt: "Link in Zelda". The results in the first row of Fig. 11 reveal that the network is optimized in the right direction but fails to reach the precise point. By contrast, the network adding the learnable triplane exhibits accurate

Figure 10. Ablation study of the diffusion guided reconstruction loss. The ToRGB module in EG3D is trained together with G_{train}. The input text is "a close-up of a woman with green hair".

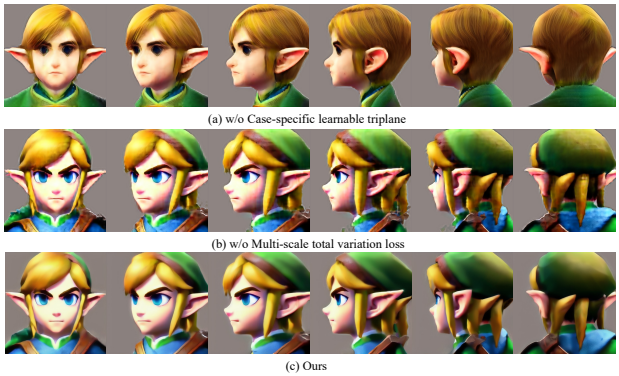


Figure 11. Ablation study toward the case-specific learnable triplane and the multi-scale total variation loss.

generation (second row in Fig. 11). Furthermore, the introduced multi-scale total variation loss L_{mstv} on the triplane facilitates more smooth results.

4.5. Applications and Limitations

Due to the page limitation, we will introduce the application of DiffusionGAN3D on real images and specify the limitations of our methods in the supplementary materials.

5. Conclusion

In this paper, we propose a novel two-stage framework DiffusionGAN3D, which boosts text-guided 3D domain adaptation and avatar generation by combining the 3D GANs and diffusion priors. Specifically, we integrate the pre-trained 3D generative models (e.g., EG3D) with the text-to-image diffusion models. The former, in our framework, set a strong foundation for text-to-avatar, enabling stable and high-quality 3D avatar generation. In return, the latter provides informative direction for 3D GANs to evolve, which facilitates the text-guided domain adaptation of 3D GANs in an efficient way. Moreover, we introduce a progressive texture refinement stage, which significantly enhances the texture quality of the generation results. Extensive experiments demonstrate that the proposed framework achieves excellent results in both domain adaptation and text-to-avatar tasks, outperforming existing methods in terms of generation quality and efficiency.

References

[1] Jonathan young. xatlas, 2021. https://triplegangers.com/. 5

[2] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3davatargan: Bridging domains for personalized editable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4552–4562, 2023. 1, 2

[3] Aibek Alanov, Vadim Titov, and Dmitry P Vetrov. Hyperdomainnet: Universal domain adaptation for generative adversarial networks. *Advances in Neural Information Processing Systems*, 35:29414–29426, 2022. 2

[4] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20950–20959, 2023. 1, 2, 5

[5] Yukang Cao, YanPei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023. 2, 3, 6

[6] 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. 2

[7] 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. 1, 2, 3, 7

[8] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 3

[9] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. *arXiv preprint arXiv:2305.02312*, 2023. 1, 2, 5

[10] Aditya Ramesh et al. Hierarchical text-conditional image generation with clip latents, 2022. 2

[11] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411. IEEE, 2017. 2

[12] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 2, 6

[13] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 2

[14] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9984–9993, 2019. 2

[15] 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. 6

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[17] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 1

[18] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 3

[19] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *arXiv preprint arXiv:2305.12529*, 2023. 3

[20] 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. 2

[21] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. *arXiv preprint arXiv:2303.17606*, 2023. 2, 3

[22] Wonjoon Jin, Nuri Ryu, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. Dr.3d: Adapting 3d gans to artistic drawings. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 2

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 3

[25] Gwanghyun Kim and Se Young Chun. Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14203–14213, 2023. 1, 2

[26] Gwanghyun Kim, Ji Ha Jang, and Se Young Chun. Podia-3d: Domain adaptation of 3d generative model across large domain gap using pose-preserved text-to-image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22603–22612, 2023. 1, 2

[27] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *arXiv preprint arXiv:2306.09329*, 2023. 2, 3, 6

[28] Yushi Lan, Xuyi Meng, Shuai Yang, Chen Change Loy, and Bo Dai. Self-supervised geometry-aware encoder for style-based 3d gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20940–20949, 2023. 2

[29] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5871–5880, 2020. 2

[30] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 2, 3, 6

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[32] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 5

[33] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[34] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pages 1–8, 2022. 2

[35] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 2

[36] Mohit Mendiratta Pan, Mohamed Elgharib, Kartik Teotia, Ayush Tewari, Vladislav Golyanik, Adam Kortylewski, Christian Theobalt, et al. Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *arXiv preprint arXiv:2306.00547*, 2023. 3

[37] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 6

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

[39] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 2

[40] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 3

[41] 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. 2

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. 2022. 2

[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. 2

[45] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2

[46] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 3

[47] Guoxian Song, Hongyi Xu, Jing Liu, Tiancheng Zhi, Yichun Shi, Jianfeng Zhang, Zihang Jiang, Jiashi Feng, Shen Sang, and Linjie Luo. Agilegan3d: Few-shot 3d portrait stylization by augmented transfer learning. *arXiv preprint arXiv:2303.14297*, 2023. 1

[48] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*, 2022. 2, 4, 6

[49] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2, 3, 8

[50] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2, 3, 6

[51] Chi Zhang, Yiwen Chen, Yijun Fu, Zhenglin Zhou, Gang Yu, Billzb Wang, Bin Fu, Tao Chen, Guosheng Lin, and Chun-

hua Shen. Styleavatar3d: Leveraging image-text diffusion models for high-fidelity 3d avatar generation. *arXiv preprint arXiv:2305.19012*, 2023. 1, 2

[52] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose. *arXiv preprint arXiv:2308.03610*, 2023. 2, 3, 6

[53] Junzhe Zhang, Yushi Lan, Shuai Yang, Fangzhou Hong, Quan Wang, Chai Kiat Yeo, Ziwei Liu, and Chen Change Loy. Deformtoon3d: Deformable neural radiance fields for 3d toonification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9144–9154, 2023. 1, 2

[54] Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance. *arXiv preprint arXiv:2304.03117*, 2023. 3

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 3, 6

[56] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 1, 2