AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control

Ruixiang Jiang
The Hong Kong Polytechnic University
rui-x.jiang@connect.polyu.hk

Can Wang
City University of Hong Kong
cwang355-c@my.cityu.edu.hk

Jingbo Zhang
City University of Hong Kong
jbzhang6-c@my.cityu.edu.hk

Menglei Chai
Google
cmlatsim@gmail.com

Mingming He
Netflix
hmm.lillian@gmail.com

Dongdong Chen
Microsoft Cloud AI
cddlyf@gmail.com

Jing Liao*
City University of Hong Kong
jingliao@cityu.edu.hk

Abstract

Neural implicit fields are powerful for representing 3D scenes and generating high-quality novel views, but it remains challenging to use such implicit representations for creating a 3D human avatar with a specific identity and artistic style that can be easily animated. Our proposed method, AvatarCraft, addresses this challenge by using diffusion models to guide the learning of geometry and texture for a neural avatar based on a single text prompt. We carefully design the optimization framework of neural implicit fields, including a coarse-to-fine multi-bounding box training strategy, shape regularization, and diffusion-based constraints, to produce high-quality geometry and texture. Additionally, we make the human avatar animatable by deforming the neural implicit field with an explicit warping field that maps the target human mesh to a template human mesh, both represented using parametric human models. This simplifies animation and reshaping of the generated avatar by controlling pose and shape parameters. Extensive experiments on various text descriptions show that AvatarCraft is effective and robust in creating human avatars and rendering novel views, poses, and shapes. Our project page is: https://avatar-craft.github.io/.

1. Introduction

Creating human avatars is crucial for content generation in various immersive media, where users can alter the character to a specific identity, apply an artistic style, or animate with simple motion control. While traditional manual authoring of digital characters often involves cumbersome and time-consuming efforts from skilled artists, the recent progress in human digitization has shown exciting potential towards more user-friendly solutions. Nevertheless, avatar generation still faces a set of tough challenges. First of all, intuitive control is highly coveted for the system to understand specific user needs in the most natural form. Second, the generated avatars should be immediately ready for applications such as view synthesis, scene composition, and retargetable animation. Finally, the avatars should be of high quality, considering both the overall visual fidelity and preservation of target styles or identities in geometry and texture, especially when being manipulated or animated.

Significant efforts have been made in search of natural user controls for avatars. One representative stream of works [7, 2, 11, 47, 52, 51, 50] takes reference images to stylize an avatar. Unfortunately, finding suitable references that perfectly match the desired shape and appearance is not always easy, substantially limiting real-world use. On the other hand, text prompts are attracting more attention as a more natural control for generating high-quality 3D avatars, with the recent advances in large-scale vision-language models. In particular, text-to-3D avatar creation [16, 56, 31, 53] is explored by leveraging the zero-shot generation ability of Contrastive Language-Image Pre-Training (CLIP) [41]. Following this trend, our work also tackles the problem of text-guided avatar creation. In particular, we aim at high-quality 3D avatar generation, which not only supports static view synthesis but also allows for controllable animation.

Text-driven avatar creation poses great challenges in pro-

[*] Corresponding Author.

ducing high-quality geometry and texture while providing flexible animation capabilities. Existing methods [31, 56, 16] address these challenges by adopting cross-modal supervision to guide the generation and modeling avatars as explicit meshes to support skeleton-driven animation [16]. However, despite the powerful representational capability of CLIP, due to the semantic structures and complex deformations of human bodies, these methods oftentimes struggle to produce detailed and consistent appearances for avatars [56]. On a parallel thread, the pioneering text-conditional diffusion models [39, 27] demonstrate stronger text-to-3D generation ability compared to CLIP. However, these approaches focus on general static objects rather than animatble human avatars. To address these challenges, for the first time ever, we propose to tackle text-guided avatar creation leveraging a diffusion model for guidance, leading to improved results in terms of consistency and quality.

Additionally, instead of mesh-based representations [31, 56], we exploit neural implicit fields [33, 54] to represent the avatar, which allow for volume rendering and generate high-quality novel views, making them especially advantageous for complex topology reconstructions and photo-realistic renderings. The implicit representation also makes it straightforward to composite the avatar with any implicit 3D scene while preserving realistic occlusions. To animate the avatar, we use SMPL to directly deform the neural implicit field, which enables a flexible way to animate and reshape the avatar by controlling the pose and shape parameters of SMPL, without requiring additional training.

In this paper, we propose *AvatarCraft*, an approach for transforming text into neural human avatars, which uses diffusion models to stylize the geometry and texture, with shape and pose controlled by parametric human models. Our method starts with a bare neural human avatar as the template. Given a text prompt, we use diffusion models [43] to guide the creation of a human avatar by updating the template with geometry and texture that are consistent with the text. However, directly applying diffusion models [43] can lead to distorted geometry and textures as the diffusion loss [39] is sensitive to the input resolution and biased towards geometry modeling. To address this issue, we design a novel coarse-to-fine multi-bounding-box training strategy, where the diffusion model guides the creation at multiple scales to improve the global style consistency while preserving fine details. We also introduce a shape regularization method to penalize the accumulated ray opacity of the avatar, to stabilize the optimization process. Regarding avatar animation, unlike skeleton-driven mesh animation in previous approaches [16], our method defines an explicit warping field that maps the target human parameterization to the template human avatar and uses the warping field to deform the neural implicit field directly. Overall, our method enables parametrized shape and motion control of avatars and supports high-quality novel view synthesis, scene composition, and animation simultaneously. In summary, our contributions are as follows:

- We present a text-guided method for creating high-quality 3D human avatars that outperform previous approaches by using diffusion models as guidance.
- Our method enables easy animation and reshaping of neural avatar radiance fields using only the pose and shape parameters of the SMPL model, without requiring any additional training.
- We demonstrate the ability to composite our neural avatar radiance fields with real neural scenes for occlusion-aware novel view synthesis.

2. Related Works

Neural Implicit Fields. To represent 3D objects, previous approaches mainly utilize explicit representations like point clouds [3, 28], voxels [10, 23], and meshes [21, 15, 11, 58]. Due to the limited representative ability, these approaches can hardly synthesize high-fidelity novel views. Recently, led by the pioneering work of NeRF [32], the advances in neural implicit fields [32, 54, 55, 33, 40, 37] have sparked research into representing 3D objects via a continuous function, producing photo-realistic novel views by volume rendering. Specifically, NeuS [54] and VolSDF [55] are proposed to improve NeRF [32] by representing the geometry using a sign distance field (SDF) instead of occupancy. Training such an implicit function takes a long time due to large amount of parameters in multi-fully connected layers. Instant-NGP [33] reduces the training cost by imposing a smaller network without sacrificing quality. This network is augmented by a multiresolution hash table of learnable feature vectors, which significantly reduces the number of floating point and memory access operations, allowing for fast training of high-quality neural graphics primitives. Therefore, to improve the geometry and reduce the training cost, we combine NeuS and Instant-NGP as our basic architecture. Although the aforementioned methods improve convergence at the cost of a little precision, they can only model static scenes. Recent works [40, 37, 48, 42] extend the static neural implicit fields to dynamic ones with an implicit deformation network to warp sampled points along the ray to deform a template. Unlike them, instead of optimizing a deformation network to learn a warping field, we deform the neural implicit fields based on the parametric human model (i.e. SMPL). We define an explicit warping field by calculating a local transformation between the source SMPL mesh and the target one while aligning the source mesh to the neural implicit fields.

Diffusion Models. Recently, diffusion models have emerged as promising and widely-attracted image generators due to their impressive generative performance [14, 36, 35, 45, 43]. In addition to directly converting Gaussian noise into images with learned data distributions through iterative denoising, these models can also generate desired images conditioned on the guidance like class labels, text, and low-resolution images [6, 43, 44, 46]. Text-driven diffusion models have demonstrated unprecedented capabilities in generating diverse high-quality semantically relevant images, such as GLIDE [35], Imagen [45], and Stable Diffusion Model (SDM) [43]. These models have been successfully applied in various domains, including image stylization, image editing, video generation, and 3D scene generation [17, 59, 12, 13, 49, 39, 27]. Among them, Diff-Styler [17] proposes a dual diffusion architecture for text-driven image stylization, which can generate text-dependent stylized images with the spatial structure consistent with the content image but cannot support novel view synthesis of stylized results due to the lack of 3D constraints. To bridge the gap between 2D images and 3D scenes, DreamFusion [39] introduces a diffusion loss to enable 3D object generation from a pretrained 2D image-text diffusion model. Furthermore, Magic3D [27] proposes a coarse-to-fine strategy for fine-grained 3D scene generation based on the pretrained diffusion model. Although DreamFusion and Magic3D enable the generation of view-consistent 3D objects from 2D prior models, 3D avatar creation remains a challenging task for them. We focus on investigating the text-guided generation of 3D human avatars with shape and pose flexibly controlled by the SMPL model.

Text-Guided 3D Avatar Creation. Aside from creating 3D objects with reference images [5, 34, 18, 1], the recent works propose to add detailed styles to a bare 3D body mesh given a text prompt as guidance. For example, Text2Mesh [31] and CLIP-Actor [56] utilize CLIP to guide the creation of a bare 3D mesh by learning a displacement map for geometry deformation and vertex colors for texture generation. Due to the limited representative ability of the mesh, such methods cannot generate detailed textures and render high-quality novel views. AvatarCLIP [16] and NeRF-Art [53] stylize a pre-trained neural implicit field guided by CLIP, producing photo-realistic renderings. AvatarCLIP first reconstructs a bare human avatar and then inpaints it using the classical CLIP similarity loss [38]. It also designs a CLIP-guided method for reference-based motion synthesis to animate the stylized 3D avatar. NeRF-Art improves the stylization results by imposing a directional CLIP loss [8] and a global-local contrastive loss. However, all these methods suffer from uneven and disordered textures generated by CLIP guidance. In addition, they lack the capability to animate a human avatar and render novel poses and views simultaneously, making them unfriendly to artists and designers. In contrast, we propose creating a 3D avatar under the guidance of diffusion models and elaborating a coarse-to-fine training strategy and shape regularization to produce visually pleasing results. Furthermore, we mitigate the animation issue by defining a local transformation between the template mesh and the target based on SMPL models, which enables easy control over the pose and shape of the stylized avatar.

3. Method

In this section, we introduce our method for creating and controlling neural implicit fields. First, we provide a preliminary explanation of our basic 3D representation in §3.1. Next, we leverage diffusion models to guide the avatar generation direction §3.2. We also propose a shape regularization approach in §3.3, as well as coarse-to-fine and multi-bbox training strategies in §3.4, to improve the generation performance. Finally, we animate and reshape the human avatar by defining an SMPL-guided deformation for the implicit fields in §3.5.

3.1. Neural Human Avatar Representation

NeuS [54] improves upon the geometry of NeRF by replacing the occupancy representation in neural implicit fields with a SDF function. NeuS is composed of a geometry function $f(\boldsymbol{x}) : \mathbb{R}^3 \mapsto \mathbb{R}^1$ and a color function $c(\boldsymbol{x}) : \mathbb{R}^3 \mapsto \mathbb{R}^3$. The geometry function $f(\boldsymbol{x})$ takes 3D positions $\boldsymbol{x}$ as input and regresses a zero-level set surface. On the other hand, the color function $c(\boldsymbol{x})$ takes $\boldsymbol{x}$ and an optional view-direction $\boldsymbol{d}$ as input, and outputs the radiance at that point. Similar to NeRF, NeuS also leverages volume rendering to achieve the pixel color of a ray $C(\boldsymbol{o}, \boldsymbol{d})$, where $\boldsymbol{o}$ and $\boldsymbol{d}$ are the origin and direction of the ray, respectively. In addition, n points are sampled $\{\boldsymbol{p}(t) = \boldsymbol{o} + t_i\boldsymbol{d} \mid i = 0, 1, ..., n\}$:

$$C(\boldsymbol{o}, \boldsymbol{d}) = \int_0^\infty w(t)c(\boldsymbol{p}(t), \boldsymbol{d})dt, \quad (1)$$

where $w(t)$ is a weighting function:

$$w(t) = \frac{\phi_s(f(\boldsymbol{p}(t)))}{\int_0^\infty f(\boldsymbol{p}(u))du}, \quad (2)$$

where ϕ_s is the logistic density distribution, which allows for better geometry representation for human avatar modeling compared to NeRF.

Training deep coordinate-based networks can be challenging due to slow performance. However, a promising solution has been proposed by Instant-NGP [33], which introduces a multi-resolution hash encoding technique to alleviate this issue. Specifically, it defines a multi-resolution voxel grid in space, as well as a table of learnable feature vectors corresponding to each voxel. To calculate the embedding at each voxel level for a query position $\boldsymbol{x}$, the feature vector of that voxel is interpolated. The positional embedding of $\boldsymbol{x}$ can be obtained by concatenating all the feature vectors. This streamlined embedding implementation

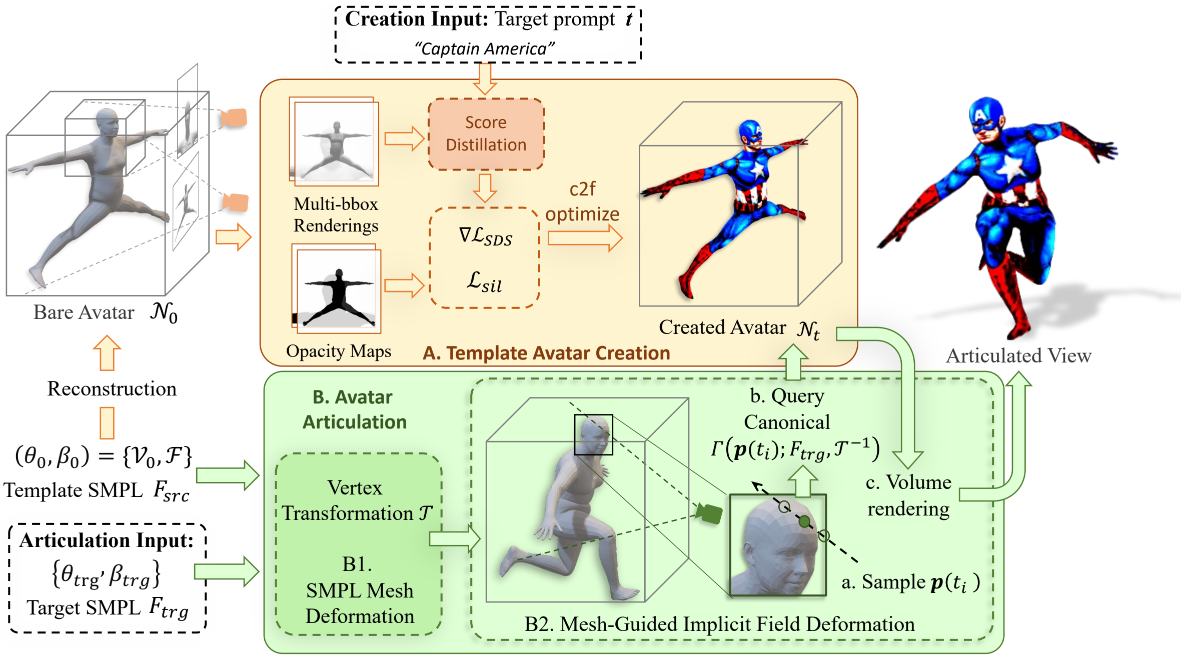
Creation Input: Target prompt t
"Captain America"
Score Distillation
Multi-bbox Renderings
$\nabla\mathcal{L}_{SDS}$
$\mathcal{L}_{sil}$
c2f optimize
Opacity Maps
A. Template Avatar Creation
Created Avatar $\mathcal{N}_t$
Bare Avatar $\mathcal{N}_0$
Articulated View
Reconstruction
$(\theta_0, \beta_0) = \{\mathcal{V}_0, \mathcal{F}\}$
Template SMPL F_{src}
Articulation Input:
$\{\theta_{trg}, \beta_{trg}\}$
Target SMPL F_{trg}
B. Avatar Articulation
Vertex Transformation $\mathcal{T}$
B1. SMPL Mesh Deformation
b. Query Canonical
$\Gamma(\boldsymbol{p}(t_i); F_{trg}, \mathcal{T}^{-1})$
c. Volume rendering
a. Sample $\boldsymbol{p}(t_i)$
B2. Mesh-Guided Implicit Field Deformation

Figure 1. **Method Overview of AvatarCraft.** The proposed pipeline is divided into two stages. A) We utilize SDS loss and additional shape regularization to create the template of target avatar using our multiple bounding-box (multi-bbox) and coarse-to-fine (c2f) training strategy. B1) We first use input SMPL parameters to calculate per-vertex rigid transformations. B2.a) Guided by F_{trg}, the camera emits rays, and points $\boldsymbol{p}(t_i)$ on the ray are sampled. B2.b) For all sampled points, we find their corresponding points in the generated canonical space $\mathcal{N}_t$ based on inverse vertex transformation $\mathcal{T}^{-1}$ as well as SMPL mesh F_{trg}. B2.c) The color of the rays can be computed using the volumetric rendering equation.

in Instant-NGP has the potential to greatly enhance the reconstruction process while ensuring minimal loss in quality. Despite its strengths, Instant-NGP has not yet achieved the same level of geometry reconstruction accuracy as NeuS.

We combine NeuS and Instant-NGP as our neural implicit model of avatars, leading to an improvement in speed while maintaining satisfactory reconstruction quality.

3.2. Diffusion-Guided Avatar Creation

We start by introducing diffusion-guided avatar creation from a text prompt. Let $\mathcal{N}_0 = \{f(\boldsymbol{x}), c(\boldsymbol{x})\}$ denote the bare avatar in canonical space, and $\boldsymbol{t}$ be a text prompt. Our objective is to optimize $\mathcal{N}_0$ such that the generated avatar $\mathcal{N}_{\boldsymbol{t}}$ is amenable to $\boldsymbol{t}$. To achieve this, we first build $\mathcal{N}_0$ by reconstructing from multi-view renderings of a bare SMPL mesh $\mathcal{M}_0 = \{\theta_0, \beta_0\} = \{\mathcal{V}_0, \mathcal{F}\}$, where θ_0 and β_0 are SMPL pose and shape parameters, and $\mathcal{V}_0$ and $\mathcal{F}$ represent vertices and faces, respectively. After reconstructing this template, we leverage the recent Score Distillation Score (SDS) loss from DreamFusion [39] to guide the creation process, which is defined as:

$$\nabla\mathcal{L}_{SDS} = \mathbb{E}_{m,\epsilon}\left[s(m)\left(\epsilon_\phi\left(z_m; m, \boldsymbol{t}\right) - \epsilon\right)\frac{\partial z_m}{\partial x}\frac{\partial x}{\partial \theta}\right], \tag{3}$$

where ϵ_ϕ is the denoiser, $s(m)$ is a weighting function depending on the timestep m, z_m is the noise, and x is the latent code encoded from the 2D rendering of the avatar. By computing the SDS loss, we enable the propagation of gradients from the diffusion model to update the neural radiance field.

3.3. Shape Regularization

Previous approaches [39, 27, 30] demonstrate the power of SDS loss in supervising 3D generation tasks. However, directly applying it to shift the shape $f(\boldsymbol{x})$ and color $c(\boldsymbol{x})$ of a human avatar to match the target style description $\boldsymbol{t}$ can yield undesired results, as shown in Fig. 3. The SDS loss is biased towards modeling geometry $f(\boldsymbol{x})$ in the early generation steps, which can result in the catastrophic destruction of the template prior $\mathcal{N}_0$ and an attempt to generate the human avatar from scratch instead. The high variance of the diffusion model in this generation process can be unstable, resulting in degenerate solutions such as empty volumes, or adversarial results such as flat geometry and multi-face issues [39, 27]. Therefore, to stabilize our generation process, we propose to regularize the shape $f(\boldsymbol{x})$ as below.

We denote the trainable parameters of the geometry and color network in neural implicit fields as $\{\Theta_f, \Theta_c\}$. A naive solution to regularize the geometry is to freeze Θ_f. How-

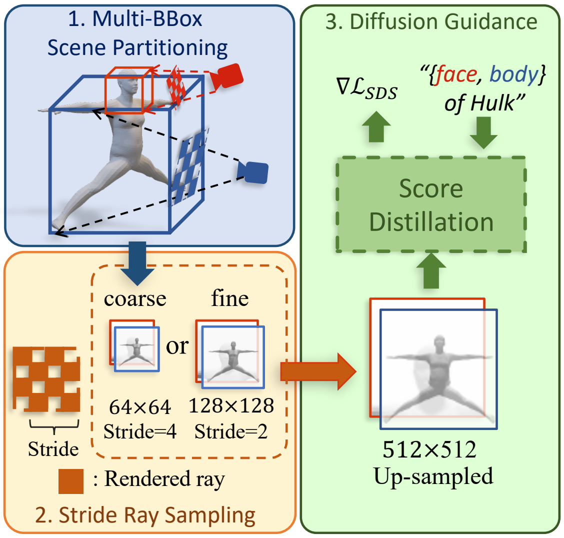

Figure 2. **Coarse-to-Fine and Multi-BBox Training.** 1) We partition the canonical space into two bounding boxes for sampling cameras. 2) we use stride ray sampling to render the avatar at different scales. 3) the rendered coarse or fine avatar is interpolated to fit stable diffusion input assumption for calculating the SDS loss.

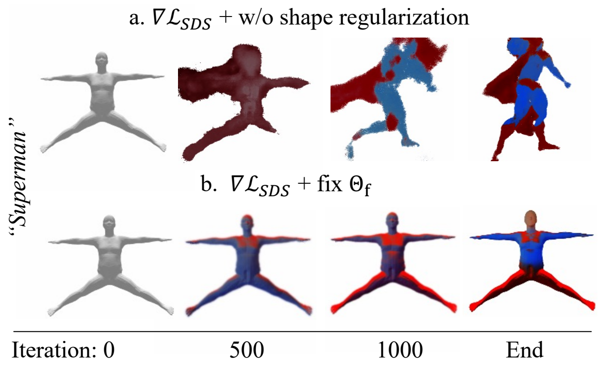
a. $\nabla\mathcal{L}_{SDS}$ + w/o shape regularization
“Superman”
b. $\nabla\mathcal{L}_{SDS}$ + fix Θ_f
Iteration: 0
500
1000
End

Figure 3. **Problems with Applying $\nabla\mathcal{L}_{SDS}$ for Avatar Generation**. We demonstrate the optimization progress of generation under two conditions: a) without any geometry constraint, and b) with the SDF parameters Θ_f fixed. Applying no constraint leads to adversarial results, while fixing the SDF parameters results in blurry texture. The prompt for both experiments is *“Superman”*.

ever, as shown in Fig. 3, this simple scheme leads to blurry and over-saturated results as it prevents the neural implicit fields from learning high-frequency details. Therefore, we aim to optimize both $\{\Theta_f, \Theta_c\}$, while regularizing the optimization of shape network. Our key idea is that the object silhouette could serve as a proxy for its geometry, while allowing for non-convex details to be sculpted [24]. Specifically, a pixel-level loss is introduced to regularize the ray opacity of the generated avatar:

$$\mathcal{L}_{sil} = \frac{1}{HW}\sum_{H,W}|O(\boldsymbol{o},\boldsymbol{d};\mathcal{N}_0) - O(\boldsymbol{o},\boldsymbol{d};\mathcal{N}_{\boldsymbol{t}})|, \tag{4}$$

where $O(\boldsymbol{o},\boldsymbol{d};\mathcal{N})$ denotes the accumulated point opacity along the ray direction $\boldsymbol{d}$:

$$O(\boldsymbol{o},\boldsymbol{d};\mathcal{N}) = \int_0^\infty w(t)dt. \tag{5}$$

Finally, the overall loss used in the generation process is:

$$\mathcal{L} = \nabla\mathcal{L}_{SDS} + \lambda_{sil}\mathcal{L}_{sil} + \lambda_{eik}\mathcal{L}_{eik}, \tag{6}$$

where $\mathcal{L}_{eik}$ is the Eikonal term [9] to regularize the normal.

3.4. Coarse-to-Fine and Multi-BBox Training

A key to improving the fidelity of a generated avatar lies in allocating the correct texture to each part of the human body. Taking the style *“Superman”* as an example, users would typically expect a large “S” on the chest and a red crotch to be the most significant details of Superman’s costume. Subsequently, users might expect additional details in the face, belt, and shoes that resemble the identity of Superman. Visual features related to these attributes span different scales, and creating an avatar at one specific scale may result in the loss or misalignment of important details. Based on this observation, we propose a coarse-to-fine generation strategy, which aims to capture style details at different scales. Specifically, we adopt a two-stage generation scheme. Firstly, we stylize $\mathcal{N}_0$ in the rendering resolution of 64×64. Then, we fine-tune it by doubling the resolution while using the same set of losses to generate texture details. The lower resolution of the initial renderings serves as a natural band-limited filter that promotes the creation of high-level textures. As the resolution increases, the generation of fine-detailed textures becomes feasible.

In addition, human perception is particularly sensitive to artifacts and distortions in facial features. However, directly stylizing the human avatar radiance fields can result in a degradation of facial features. To address this limitation, we devise a solution to divide the scene into face and body bounding boxes according to the SMPL prior $\mathcal{M}_0$. Our approach involves dedicating the face box exclusively to rendering the head and neck, while the body box is used to render the entire avatar, including the head. Besides, we also augment the prompt $\boldsymbol{t}$ according to the rendered box. This approach enhances the fidelity of facial features while maintaining a natural transition between the head and the body. By employing this coarse-to-fine and multi-bounding-box training strategy, we can improve the alignment of visual features across the entire avatar body.

3.5. SMPL-Guided Avatar Articulation

Avatar $\mathcal{N}_{\boldsymbol{t}}$ in the canonical space is intrinsically aligned with the underlying SMPL model $F_{src} = \{\theta_0, \beta_0\} = \{\mathcal{V}_0, \mathcal{F}\}$. Given the target model $F_{trg} = \{\theta_{trg}, \beta_{trg}\} =$

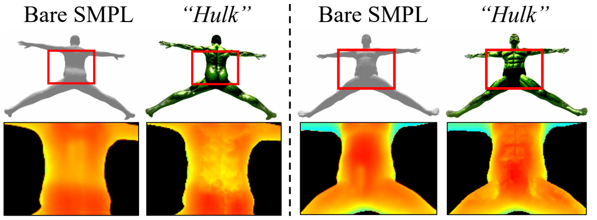

Figure 4. **Geometry Generation.** AvatarCraft could generate fine geometry detail on the avatar surface. We show the rendered depth map of bare SMPL and *"hulk"*.

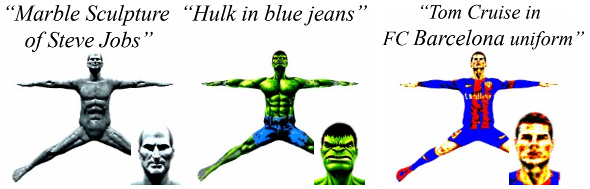
"Marble Sculpture of Steve Jobs"
"Hulk in blue jeans"
"Tom Cruise in FC Barcelona uniform"

Figure 5. **Concept Mixing.** AvatarCraft could generate novel avatars by mixing different concepts together.

$\{\mathcal{V}_{trg}, \mathcal{F}\}$, our method aims to deform the template avatar to render novel views with the target pose and shape. We demonstrate that our approach provides an intuitive way to control both the shape and the pose of the implicit avatar without the need for training, enabling various applications such as avatar shape customization and animation.

SMPL Mesh Deformation. We first establish the correspondence between F_{src} and F_{trg}. Specifically, for each vertex, we use beta-blended shape [29] to compute the shape-related vertex deformation $\mathcal{T}^{\beta_{trg}} \in \mathrm{SE}(3)$ from $\{\theta_0, \beta_0\}$ to $\{\theta_0, \beta_{trg}\}$. Next, we calculate the shape-related vertex rigid transformations $\mathcal{T}^{\theta_{trg}} \in \mathrm{SE}(3)$ from $\{\theta_0, \beta_{trg}\}$ to $\{\theta_{trg}, \beta_{trg}\}$ using Linear Blended Skining (LBS) algorithm. Since SMPL mesh topology is agnostic to shape and pose by definition, we can obtain a bijective mapping of vertices between the two meshes as $\mathcal{T} = \mathcal{T}^{\theta_{trg}}\mathcal{T}^{\beta_{trg}}$.

Mesh-Guided Implicit Field Deformation. After obtaining the vertex mappings $\mathcal{T}$, we warp the positions of the sampled points to render novel views with the target shape and pose. We rewrite the volume rendering as:

$$C(\boldsymbol{o}, \boldsymbol{d}) = \int_0^\infty \eta(\boldsymbol{p}(t)) \left[w(t) c(\Gamma(\boldsymbol{p}(t); F_{trg}, \mathcal{T}^{-1})) \right] dt, \tag{7}$$

where $\Gamma(\boldsymbol{x}; F_{trg}, \mathcal{T}^{-1})$ is a transfer function that maps points $\boldsymbol{x}$ from the observation space to the canonical space, guided by the target mesh and the inverse vertex $\mathcal{T}^{-1}$ [20, 57]. To mitigate cloudy artifacts due to inaccurate warping, we adopt a mask function $\eta(\boldsymbol{p}(t))$ [4] to set the density of the points far from the target mesh surface to zero.

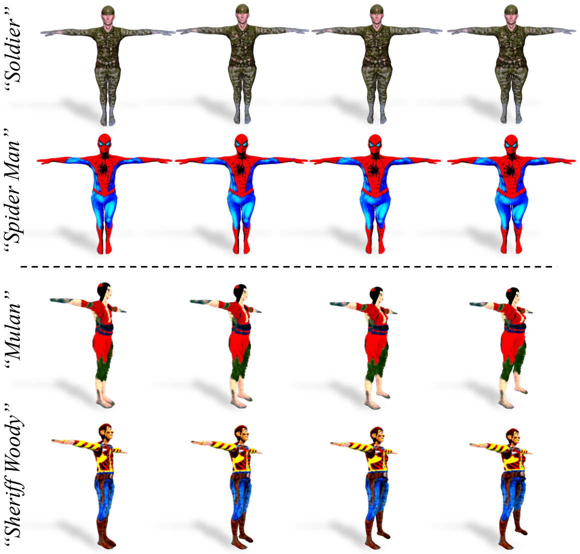

Figure 6. **Shape Interpolation.** AvatarCraft can reshape the generated avatar without the need for retraining. We demonstrate two different interpolation sequences of shape β across four styles.

4. Experiment Result and Analysis

4.1. Implementation Details

In terms of network structure, we utilize the default setting for the hash-grid [60]. Additionally, the SDF network has a depth of 2, while the color network has a depth of 3. We adopt similar augmentation strategies to those used in recent text-to-3D generative models [16, 39, 25, 19, 27]. Specifically, these include: (1) random camera extrinsic augmentation, (2) random background augmentation, and (3) view-dependent prompt augmentation. Further details about each of these augmentations can be found in the Appendix.

We reconstruct the bare SMPL model at a resolution of 512×512 using the hyper-parameters suggested by [60] for $40,000$ steps, which takes approximately 20 minutes. For avatar creation, the coarse training involves 40 epochs with 100 body captures and 20 head captures for each epoch, while the fine training involves 10 epochs with 100 body captures and 50 head captures. We train both stages using the Adam [22] optimizer with a learning rate of $5e-3$. The coarse stage takes around 30 minutes, while the fine stage takes 100 minutes. For Score Distillation, we use pretrained Stable Diffusion [43] v1.5 model, with a guidance scale of 100 for all prompts. We conduct experiments on one NVIDIA A100 GPU.

4.2. Qualitative Result

In this section, we provide qualitative results of AvatarCraft. We will show that the proposed method is effective

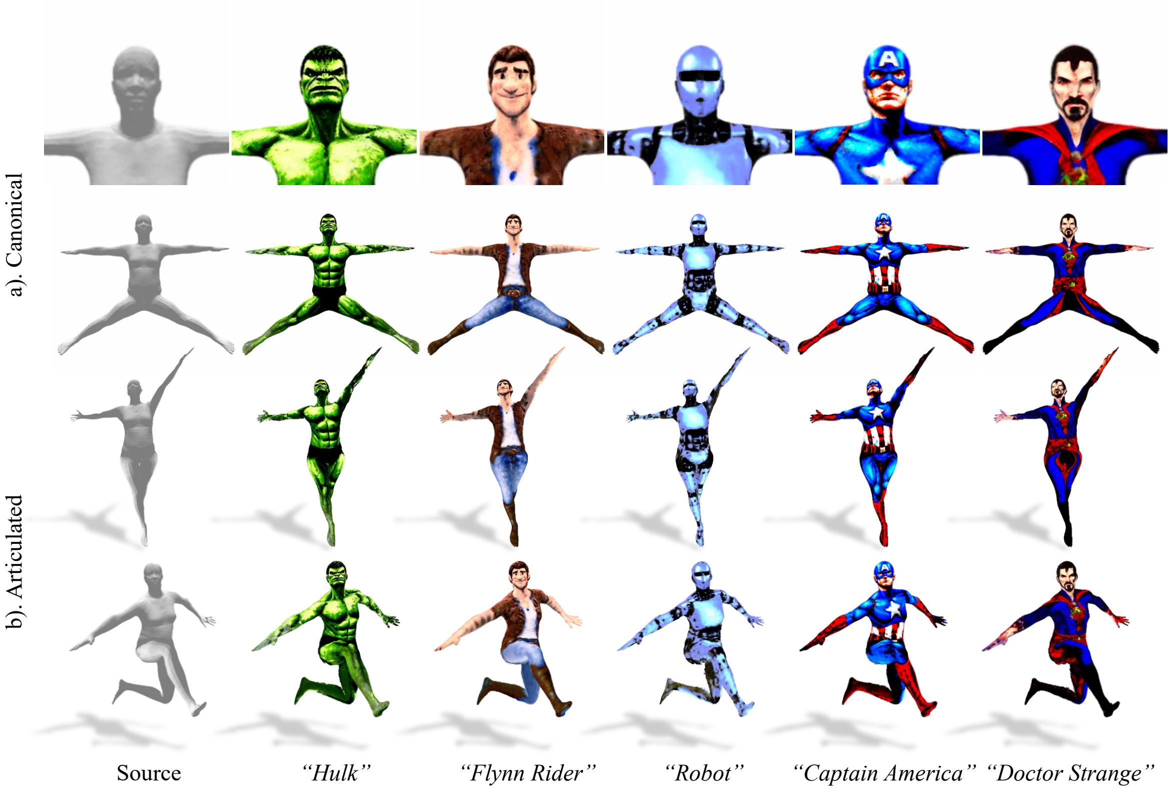

Figure 7. **Qualitative Result.** a.) AvatarCraft generates intricate details that is consistent and aligned with input prompt. b). The generated avatar could be articulated by shape and pose parameters.

for appearance generation, geometry generation, concept mixing, as well as parametric articulation.

Appearance Generation. As shown in the Fig. 7-a, AvatarCraft faithfully generates the distinctive features of representative identities, such as the muscular body of the character *"Hulk"*, the nose of *"Flynn Rider"*, and the costume of *"Captain America"*. This demonstrates the ability of AvatarCraft to produce textures that are semantically consistent and aligned with the target text descriptions. Additionally, AvatarCraft is capable of generating high-fidelity details not only on the human body but also on the head. For example the machine crew of *"Robot"* and the red cape and black beard of *"Doctor Strange"*.

Geometry Generation. In addition to appearance generation, AvatarCraft is also capable of carving geometry details on the avatar body. For instance, in Fig. 4 we show the rendered depth map of original bare SMPL and avatar *"Hulk"*. We observe clear muscle structure being formed that is consistent with the generated texture.

Concept Mixing. A strength of text-to-image models is their creativity in generating images that mix different concepts together using text guidance. Thanks to the diffusion constraints, our AvatarCraft can generate novel avatars (i.e., to dream avatars [16, 39]) by mixing different identities together. This would allow more fine-grained control over avatar style, clothing, and face. We provide examples for those cases in Fig. 5. As the result shows, our proposed method successfully generates avatars that match the input prompt.

Parametric Articulation. As an articulated-NeRF based method, AvatarCraft can render novel views with parametric control over poses and shapes simultaneously, while preserving render quality. This is achieved by defining an explicit warping field, represented by a local transformation between the source and target SMPL meshes. By referring to the generated avatar in canonical space, this warping operation enables rendering of avatar from novel views, poses, and shapes, without the need of re-training. We show shape and pose articulation in Fig. 6 and Fig. 7-b, respectively.

4.3. Comparisons

We compare AvatarCraft with state-of-the-art methods for text-driven human avatar creation, including the mesh based method of CLIP-Actor [56] and implicit-field based methods of AvatarCLIP [16] and NeRF-Art [53]. We train each method using the configurations suggested by its respective authors and present results in Fig. 8.

Compared to existing methods, AvatarCraft stands out for the high level of detail in both the avatar body and head,

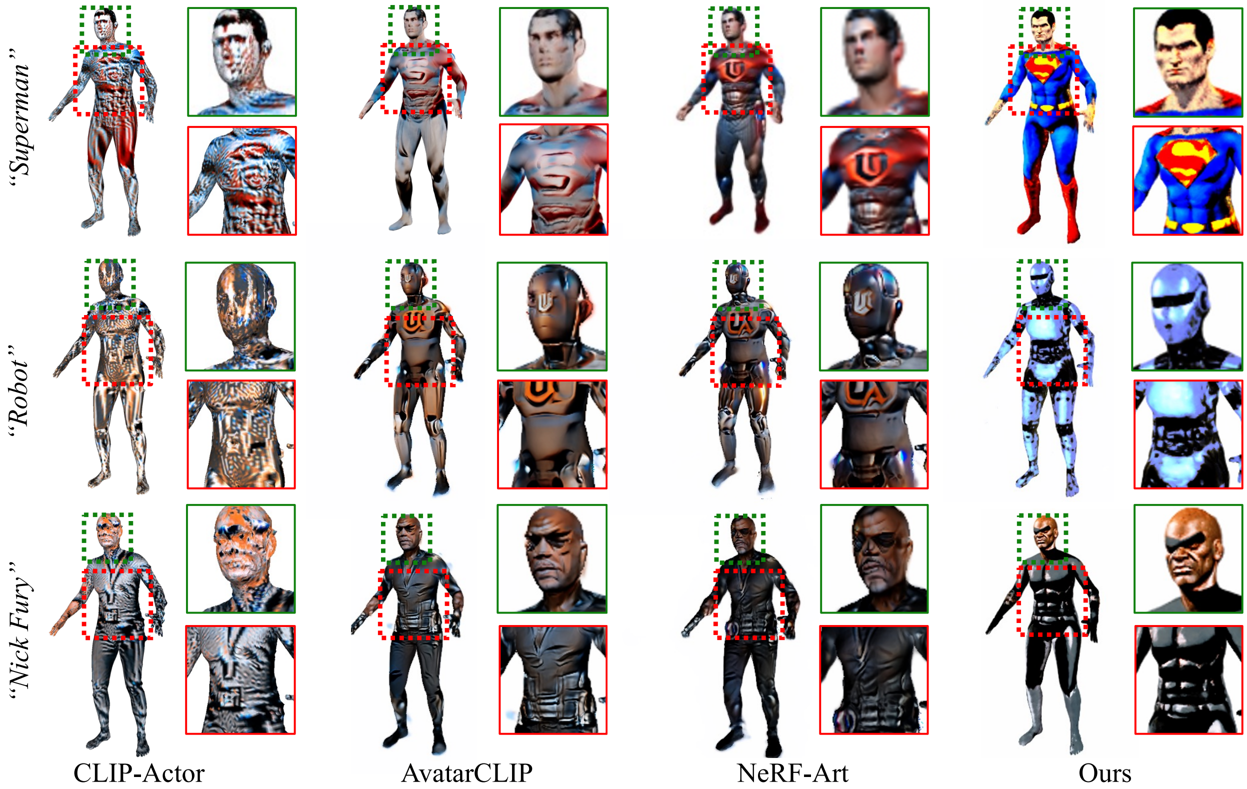

Figure 8. **Comparisons**. We compare AvatarCraft with state-of-the-art avatar creation methods, including CLIP-Actor [56], AvatarCLIP [16], and NeRF-Art [53]. Our model achieves better results both globally and locally.

attributed to its diffusion constraint and our coarse-to-fine and multi-bbox training strategy. For example, it faithfully generates intricate details like the distinctive "S" pattern on *"Superman"*'s chest that other works struggle to capture with equal quality. This demonstrates the strength of AvatarCraft in generating high quality body textures. In addition, our method produces finely detailed faces, such as the eye patch on *"Nick Fury"*, while other works only present rough facial features.

Another advantage of AvatarCraft is its ability to generate balanced textures for human avatars. This is achieved through the incorporation of a diffusion constraint, which employs a stronger and more advanced language-vision model. In contrast, other works tend to produce unreasonable and uneven textures. We postulate that it may be due to CLIP constraint that perceives and embeds the image as a whole, while being weak in providing location-aware supervision [61, 26] that is necessary for allocating correct texture locally. In contrast, the diffusion constraint employs a denoising process to provide supervision signals on the pixel level, enabling more localized guidance. Therefore, it could result in more balanced and detailed textures for avatar generation.

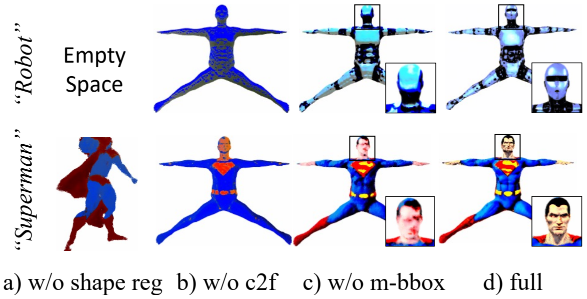

Figure 9. **Ablation Study.** We show results: (a) without the shape regularization; (b) without coarse-to-fine (c2f) training; (c) without multi-bbox training; and (d) our full method.

4.4. Ablation Analysis

To evaluate the design choices of AvatarCraft, we conduct an ablation study on the effectiveness of 1) the shape regularization, 2) the coarse-to-fine training strategy, and 3) the multi-bbox training strategy. We removed each component from AvatarCraft and reported the results in Fig. 9. For 2), we train the model at a resolution of 128 × 128 for 20 epochs. For 3), we train the model only with the body bounding box for the same number of epochs.

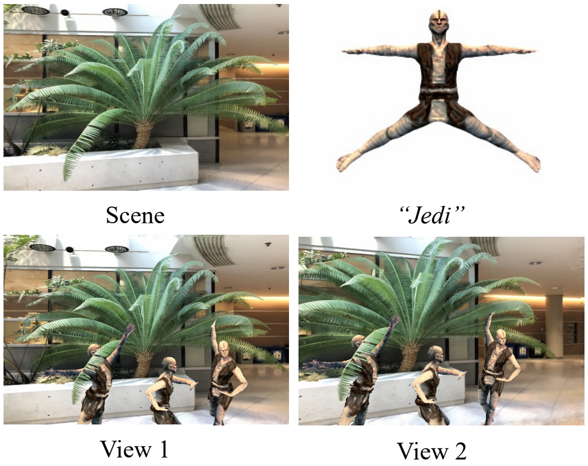

Figure 10. **Composite Rendering.** AvatarCraft enables occlusion-aware composite renderings of the created avatar with real neural scenes. We show key frames of *"Jedi"* dancing in the scene.

Shape regularization is a crucial component of AvatarCraft, as it ensures that the generated avatars have anatomically plausible shapes and proportions. By encouraging the model to learn representations consistent with prior knowledge of human anatomy, shape regularization leads to more accurate avatar creations. Without shape regularization, the method may produce unreasonable or even failed results.

The coarse-to-fine training enables alignment of fine-grained texture generation both globally and locally. In contrast, methods that do not use coarse-to-fine training only produce coarse colors with no semantically meaningful local details. Our multi-bbox training technique enhances the fidelity of facial features while maintaining a natural transition between the head and body. The methods without multi-bbox training may degrade facial features.

4.5. Application: Composite Rendering

Our method enables composite rendering of the avatar along with realistic neural scenes. To achieve this, we manually align the avatar with the scene and render them with the same camera. However, as our avatar is represented as a SDF and implicit scenes are usually encoded as occupancy fields (e.g., NeRF), we cannot directly apply Eq. 1 to a ray that contains points sampled from both representations. To circumvent this challenge, we primarily perform a depth test to achieve occlusion-aware composite rendering as illustrated in Fig. 10.

5. Conclusion

We present AvatarCraft, a novel approach to generating human avatars using text guidance and an implicit neural representation with parameterized shape and pose control. Unlike existing methods that struggle to produce visually-pleasing results when using CLIP guidance, our approach utilizes diffusion models to match the desired text description and generate fine-grained details in both geometry and texture. Furthermore, our method allows for easy animation and reshaping of the avatar using SMPL parameters and requires no training for novel pose and shape synthesis. Additionally, the use of neural implicit fields provides advantages over mesh-based avatars in terms of photorealistic rendering and easy composition with implicit 3D scenes.

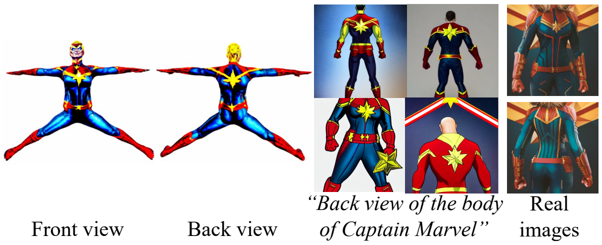

Figure 11. **Limitation.** We show the created avatar, 2D images generated by the base diffusion model, and reference real images.

Limitation. Fig. 11 demonstrates one limitation of our approach, where the diffusion model has difficulty generating textures of equal quality on the back of the avatar due to its limited ability to conceptualize views like the back that were underrepresented in the training data. Consequently, the generated back textures may exhibit certain inconsistencies compared to the reference ground-truth. To address this issue, we plan to explore the possibility of training the diffusion model with more diverse human data to enhance its ability to generate faithful and accurate textures.

Acknoledgement

This work was supported by a GRF grant (Project No. CityU 11208123) from the Research Grants Council (RGC) of Hong Kong.

References

[1] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3davatargan: Bridging domains for personalized editable avatars. *arXiv preprint arXiv:2301.02700*, 2023. 3

[2] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 1

[3] Xu Cao, Weimin Wang, Katashi Nagao, and Ryosuke Nakamura. Psnet: A style transfer network for point cloud stylization on geometry and color. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3337–3345, 2020. 2

[4] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural

radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629*, 2021. 6

[5] Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. Stylizing 3d scene via implicit representation and hypernetwork. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1475–1484, 2022. 3

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[7] Jakub Fišer, Ondřej Jamriška, Michal Lukáč, Eli Shechtman, Paul Asente, Jingwan Lu, and Daniel Sỳkora. Stylit: illumination-guided example-based stylization of 3d renderings. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 1

[8] 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. 3

[9] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5

[10] Jie Guo, Mengtian Li, Zijing Zong, Yuntao Liu, Jingwu He, Yanwen Guo, and Ling-Qi Yan. Volumetric appearance stylization with stylizing kernel prediction network. *ACM Trans Graph*, 40:1–15, 2021. 2

[11] Fangzhou Han, Shuquan Ye, Mingming He, Menglei Chai, and Jing Liao. Exemplar-based 3d portrait stylization. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1371–1383, 2021. 1, 2

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[13] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[15] Lukas Höllein, Justin Johnson, and Matthias Nießner. Stylemesh: Style transfer for indoor 3d scene reconstructions. *arXiv preprint arXiv:2112.01530*, 2021. 2

[16] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 1, 2, 3, 6, 7, 8

[17] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Yong Zhang, Weiming Dong, and Changsheng Xu. Diffstyler: Controllable dual diffusion for text-driven image stylization. *arXiv preprint arXiv:2211.10682*, 2022. 3

[18] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18342–18352, 2022. 3

[19] 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. 6

[20] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 402–418. Springer, 2022. 6

[21] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 2

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Oliver Klehm, Ivo Ihrke, Hans-Peter Seidel, and Elmar Eisemann. Property and lighting manipulations for static volume stylization using a painting metaphor. *IEEE Transactions on Visualization and Computer Graphics*, 20(7):983–995, 2014. 2

[24] Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162, 1994. 5

[25] Han-Hung Lee and Angel X Chang. Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*, 2022. 6

[26] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 8

[27] 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. *arXiv preprint arXiv:2211.10440*, 2022. 2, 3, 4, 6

[28] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 6

[30] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 4

[31] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022. 1, 2, 3

[32] 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. In *European conference on computer vision*, pages 405–421. Springer, 2020. 2

[33] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2, 3

[34] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *arXiv preprint arXiv:2207.02363*, 2022. 3

[35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[36] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[37] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[38] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[39] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3, 4, 6, 7

[40] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[41] 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. 1

[42] Jian Ren, Menglei Chai, Oliver J Woodford, Kyle Olszewski, and Sergey Tulyakov. Flow guided transformable bottleneck networks for motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10795–10805, 2021. 2

[43] 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, 3, 6

[44] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[46] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[47] Shen Sang, Tiancheng Zhi, Guoxian Song, Minghao Liu, Chunpong Lai, Jing Liu, Xiang Wen, James Davis, and Linjie Luo. Agileavatar: Stylized 3d avatar creation via cascaded domain bridging. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 1

[48] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Hsin-Ying Lee, Jian Ren, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. *arXiv preprint arXiv:2301.11326*, 2023. 2

[49] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[50] Aneta Texler, Ondřej Texler, Michal Kučera, Menglei Chai, and Daniel Sỳkora. Faceblit: instant real-time example-based style transfer to facial videos. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 4(1):1–17, 2021. 1

[51] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclei Chai, Sergey Tulyakov, and Daniel Sỳkora. Interactive video stylization using few-shot patch-based training. *ACM Transactions on Graphics (TOG)*, 39(4):73–1, 2020. 1

[52] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Cross-domain and disentangled face manipulation with 3d guidance. *arXiv preprint arXiv:2104.11228*, 2021. 1

[53] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *arXiv preprint arXiv:2212.08070*, 2022. 1, 3, 7, 8

[54] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2, 3

[55] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2

[56] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 173–191. Springer, 2022. 1, 2, 3, 7, 8

[57] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of

neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18353–18364, 2022. 6

[58] Mohan Zhang, Jing Liao, and Jinhui Yu. Deep exemplar-based color transfer for 3d model. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 2

[59] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. *arXiv preprint arXiv:2212.04489*, 2022. 3

[60] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, et al. Human performance modeling and rendering via neural animated mesh. *arXiv preprint arXiv:2209.08468*, 2022. 6

[61] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 8