CharacterGen: Efficient 3D Character Generation from Single Images with Multi-View Pose Canonicalization

HAO-YANG PENG, BNRist, Department of Computer Science and Technology, Tsinghua University, China
JIA-PENG ZHANG, Zhili College, Tsinghua University, China
MENG-HAO GUO, BNRist, Department of Computer Science and Technology, Tsinghua University, China
YAN-PEI CAO, VAST, China
SHI-MIN HU, BNRist, Department of Computer Science and Technology, Tsinghua University, China

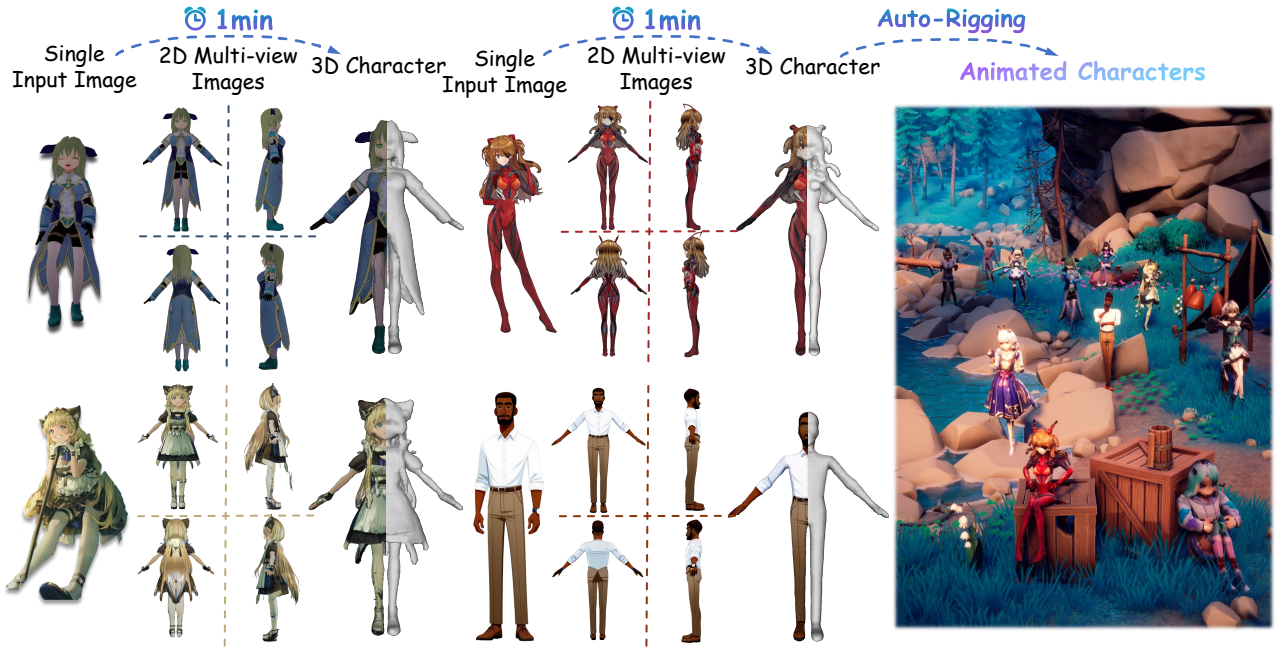

Fig. 1. CharacterGen is an efficient 3D character generation framework. It takes a single input image and generates high-quality 3D character mesh in a canonical pose with consistent appearance, suitable for downstream rigging and animation workflows. © kinoko7

In the field of digital content creation, generating high-quality 3D characters from single images is challenging, especially given the complexities of various body poses and the issues of self-occlusion and pose ambiguity. In this paper, we present CharacterGen, a framework developed to efficiently generate 3D characters. CharacterGen introduces a streamlined generation pipeline along with an image-conditioned multi-view diffusion model. This model effectively calibrates input poses to a canonical form while retaining key attributes of the input image, thereby addressing the challenges posed by diverse poses. A transformer-based, generalizable sparse-view reconstruction model is the other core component of our approach, facilitating the creation of detailed 3D models from multi-view images. We also adopt a texture-back-projection strategy to produce high-quality texture maps. Additionally, we have curated a dataset of anime characters, rendered in multiple poses and views, to train and evaluate our model. Our approach has been thoroughly evaluated through quantitative and qualitative experiments, showing its proficiency in generating 3D characters with high-quality shapes and textures, ready for downstream applications such as rigging and animation.

Authors' Contact Information: Hao-Yang Peng, phy22@mails.tsinghua.edu.cn, BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing, China; Jia-Peng Zhang, zhangjp20@mails.tsinghua.edu.cn, Zhili College, Tsinghua University, Beijing, China; Meng-Hao Guo, gmh20@mails.tsinghua.edu.cn, BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing, China; Yan-Pei Cao, caoyanpei@gmail.com, VAST, Beijing, China; Shi-Min Hu, shimin@tsinghua.edu.cn, BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing, China.

ACM 1557-7368/2024/7-ART84
https://doi.org/10.1145/3658217

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Shape modeling**; **Image manipulation**.

Additional Key Words and Phrases: Image-Driven Generation, 3D Avatar Generation, Avatar Pose Canonicalization

ACM Reference Format:
Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. 2024. CharacterGen: Efficient 3D Character Generation from Single Images with Multi-View Pose Canonicalization. *ACM Trans. Graph.* 43, 4, Article 84 (July 2024), 13 pages. https://doi.org/10.1145/3658217

1 INTRODUCTION

The digital content industry's rapid evolution has made the creation of high-quality 3D content a pivotal aspect across various domains, including film, video gaming, online streaming, and virtual reality (VR). Although manually modeled 3D content can attain exceptional quality, the significant time and labor investment required presents a substantial bottleneck. Addressing this, there has been a notable influx of exciting research [Alwala et al. 2022; Deng et al. 2023; Jun and Nichol 2023; Liu et al. 2023b,d; Long et al. 2023; Melas-Kyriazi et al. 2023; Nichol et al. 2022; Qian et al. 2023; Raj et al. 2023; Tang et al. 2023a; Wang et al. 2021b; Wang and Shi 2023; Wen et al. 2023; Wu et al. 2023a; Yoo et al. 2023] focused on generating 3D models from single images. This approach substantially lowers the barrier to entry for novice users, democratizing access to 3D content creation and potentially revolutionizing the field.

3D character models often feature complex articulations, leading to frequent self-occlusion in 2D images that pose significant challenges in reconstruction, generation, and animation. Moreover, these characters may assume a range of body poses, including some that are rare and challenging to accurately interpret, leading to a diverse yet imbalanced data domain. These complexities hinder the effective generation, rigging, and animating of such models. As a result, general 3D generation techniques [Chen et al. 2023a; Lin et al. 2023a; Metzer et al. 2023; Poole et al. 2023; Wang et al. 2023a,b] and single-view 3D reconstruction methods [Alwala et al. 2022; Hong et al. 2023b; Liu et al. 2023d; Thai et al. 2021; Wang et al. 2021b; Wen et al. 2023; Wu et al. 2023a] often fall short in delivering optimal outcomes. Prior research [Cao et al. 2023; Huang et al. 2023a,b; Kolotouros et al. 2023; Liao et al. 2023; Saito et al. 2019, 2020; Wang et al. 2023c; Xiu et al. 2023, 2022; Zheng et al. 2022] has explored the use of parametric models of 3D human bodies [Alldieck et al. 2021; Li et al. 2017; Loper et al. 2015; Pavlakos et al. 2019] as 3D priors. However, these methods are predominantly tailored to realistic human proportions and relatively tight clothing, limiting their applicability. This constraint is especially noticeable in the context of stylized characters, known for their exaggerated body proportions and complex clothing designs, which challenge the adaptability and effectiveness of these approaches.

In this paper, we introduce CharacterGen, a new approach for 3D character generation in a canonical pose from a single image. Our method stands out significantly from previous ones by allowing any input body pose in the input image and outputting a clean 3D character model. The foundational principle of CharacterGen hinges on simultaneously canonicalizing body poses and producing consistent multi-view images during the generation process. This is achieved by transforming each pose into a canonical "A-pose", a stance widely utilized in 3D character modeling, while concurrently ensuring image consistency across multiple views. This dual approach effectively addresses the challenges of self-occlusion and ambiguous human poses, significantly streamlining subsequent reconstruction, rigging, and animation stages.

Our 3D character generation approach has two tightly interconnected stages: initially, lifting a single image to multiple viewpoints while simultaneously canonicalizing the input pose; following this, we proceed to reconstruct a 3D character using this canonical pose. This method is supported by *two key insights*: *firstly*, it incorporates established principles and successful techniques from recent advancements in controllable image generation [Ye et al. 2023; Zhang and Agrawala 2023]; *secondly*, it overcomes the challenges associated with sparse-view reconstruction for 3D characters. By focusing on the canonical pose, in which the geometric and texture structures are more clearly defined and self-occlusion is minimized, our approach simplifies the task of reconstructing both geometry and texture from limited views. The first stage involves a diffusion-based, image-conditioned multi-view generation model [Liu et al. 2023a; Shi et al. 2023a; Tang et al. 2023b; Wang and Shi 2023], adept at capturing and translating both the global and local character features from the input image to the canonical pose, which further facilitates the generation of consistent canonical pose images across multiple views. The second stage employs a transformer-based, generalizable sparse-view reconstruction model [Hong et al. 2023b]. This model is key to generating a coarsely textured 3D character model from the images produced in the first stage. We further refine the model's texture resolution through projective texture mapping and Poisson Blending [Pérez et al. 2003], achieving a detailed final model. Furthermore, generating characters in a canonical pose also significantly benefits downstream applications, such as rigging and animation. We show the generated 3D characters with animation in Fig. 1. Our whole generation process takes less than 1 minute.

To train our pipeline, we have compiled a multi-pose, multi-view character dataset, focusing on anime characters due to their widespread availability online, notably on platforms such as VRoid Hub [1]. We have amassed a collection of 13,746 characters and rendered these from various viewpoints across multiple body poses. This extensive collection has been organized into a dataset that we refer to as Anime3D.

In summary, our paper makes the following key contributions:

- An image-conditioned diffusion model that effectively generates multi-view consistent images of characters in a controlled canonical pose from varying input poses, addressing challenges such as self-occlusion and pose ambiguity.
- A streamlined pipeline combining our diffusion model for multi-view image generation and a transformer-based reconstruction model. This pipeline efficiently transforms single-view inputs into detailed 3D character models.
- A curated dataset of 13,746 anime characters, rendered in multiple poses and views, providing a diverse training and evaluation resource for our model and future research in 3D character generation.

2 RELATED WORKS

This section mainly discusses related works on diffusion-based 3D objects and avatar generation. Our CharacterGen also adopts

[1]

transformer-based reconstruction models [Hong et al. 2023b; Li et al. 2023] for efficient 3D character generation. Space limitations preclude discussion of 3D human reconstruction works [Cha et al. 2023; Saito et al. 2019, 2020; Xiu et al. 2023, 2022].

2.1 Diffusion-Based 3D Object Generation

Recently, diffusion methods have shown a strong ability to guide 3D object generation tasks in the past year. Pioneering work of DreamFusion [Poole et al. 2023] and SJC [Wang et al. 2023a] utilize score distillation sampling (SDS) to provide gradient guidance from pre-trained 2D diffusion models for text-to-3D generation tasks. Magic3D [Lin et al. 2023a] and Fantasia3D [Chen et al. 2023a] utilize an implicit tetrahedral field to support rendering with high resolution in the refinement stage. ProlificDreamer [Wang et al. 2023b] proposes VSD to distill gradient scores from a LoRA network to better learn the distribution of 3D objects. Zero123 [Liu et al. 2023c] presents a novel diffusion model to generate multi-view images that conform to the input image with given camera poses. Magic123 [Qian et al. 2023] combines both SDS and Zero123 guidance for generating 3D objects from image prompts, and adopts reconstruction loss to enhance front-view texture quality. MVDream [Shi et al. 2023b] and ImageDream [Wang and Shi 2023] utilize multi-view diffusion models to provide highly consistent guidance in 3D object generation process. SyncDreamer [Liu et al. 2023a] utilizes 3D-aware attention modules to achieve synchronized multi-view image generation. Various other works [Erkoç et al. 2023; Hui et al. 2022; Jun and Nichol 2023; Müller et al. 2023; Nichol et al. 2022; Zeng et al. 2022; Zhang et al. 2023b] employ 3D data to train diffusion models for direct 3D object generation, which are fast, but struggle with output diversity.

2.2 3D Avatar Generation

Using strong human-body priors such as SMPL [Loper et al. 2015] and SMPL-X [Pavlakos et al. 2019], it is possible to generate high-quality human avatars based on the general 3D generation methods. EVA3D [Hong et al. 2023a] combines a GAN backbone with a pose-guided sampling method to generate high-quality 3D human avatars. AvatarCLIP [Hong et al. 2022] first solves text-to-human generation by utilizing the pre-trained CLIP model to guide optimization of geometry and color networks.

Dreamavatar [Cao et al. 2023] and AvatarCraft [Jiang et al. 2023] utilize SMPL to initialize the implicit human geometry used in the diffusion-guided generation process. DreamHuman [Kolotouros et al. 2023] adopts ImGHum [Alldieck et al. 2021] as body priors and proposes a focus rendering mechanism to better reconstruct the detailed geometry of avatars. DreamWaltz [Huang et al. 2023a] utilizes ControlNet [Zhang and Agrawala 2023] to provide pose guidance to finetuning the animated representation. AvatarVerse [Zhang et al. 2023a] and AvatarStudio [Zhang et al. 2023c] both utilize a DensePose-guided ControlNet in the generation process to circumvent the multi-face "Janus" problem and to support part geometry optimization. TeCH [Huang et al. 2023b] supports image-prompt avatar generation by training an additional DreamBooth model [Ruiz et al. 2023] on the input image for the SDS guidance model. TADA [Liao et al. 2023] directly distills 2D diffusion models to optimize the normals and displacements of an SMPL body mesh.

Fig. 2. An example character from our Anime3D dataset from four different camera views, demonstrates how we organize the image pairs during training to extend UNet's ability to determine a canonical pose.

Most of these methods mainly focus on text-to-3D character generation and cannot utilize image prompts, which is necessary for controllable character generation. The DreamBooth-based methods are hampered severely by the "Janus" problem due to the strong front-view biases caused by overfitting on the single input image.

3 METHOD

We now explain the overall framework of our CharacterGen, which aims to efficiently generate A-pose 3D characters from 2D images in arbitrary poses. Sec. 3.1 first introduces our Anime3D dataset to show how we organize our data to assist the diffusion models in 3D spatial understanding and character pose canonicalization. Sec. 3.2 then explains how CharacterGen generates highly consistent multi-view pose-canonicalized character images. Finally, Sec. 3.3 shows our efficient 3D reconstruction pipeline.

3.1 Anime3D Dataset

To further improve diffusion models' ability to understand 3D characters and to alleviate the "Janus" problem, we have prepared the Anime3D dataset with 13,746 stylized character subjects.

3.1.1 Data Acquisition. Existing large 3D object datasets like Objaverse [Deitke et al. 2023] or OmniObject3D [Wu et al. 2023b] do not contain enough 3D stylized characters for our training purposes. Inspired by PAniC3D [Chen et al. 2023b], we first collected a large dataset of nearly 14,500 anime characters from the VRoid-Hub [VRoid 2022] and then removed non-humanoid data, to leave 13,746 character models.

3.1.2 Data Processing. We need to render all the objects into image format to fine-tune 2D diffusion models. We utilize the three-vrm [Pixiv 2019] framework to render these characters.

We first obtain "A-pose" characters and posed characters to generate canonical pose and random pose image pairs. For A-pose characters, we set the joint rotation of the left and right arm to 45°

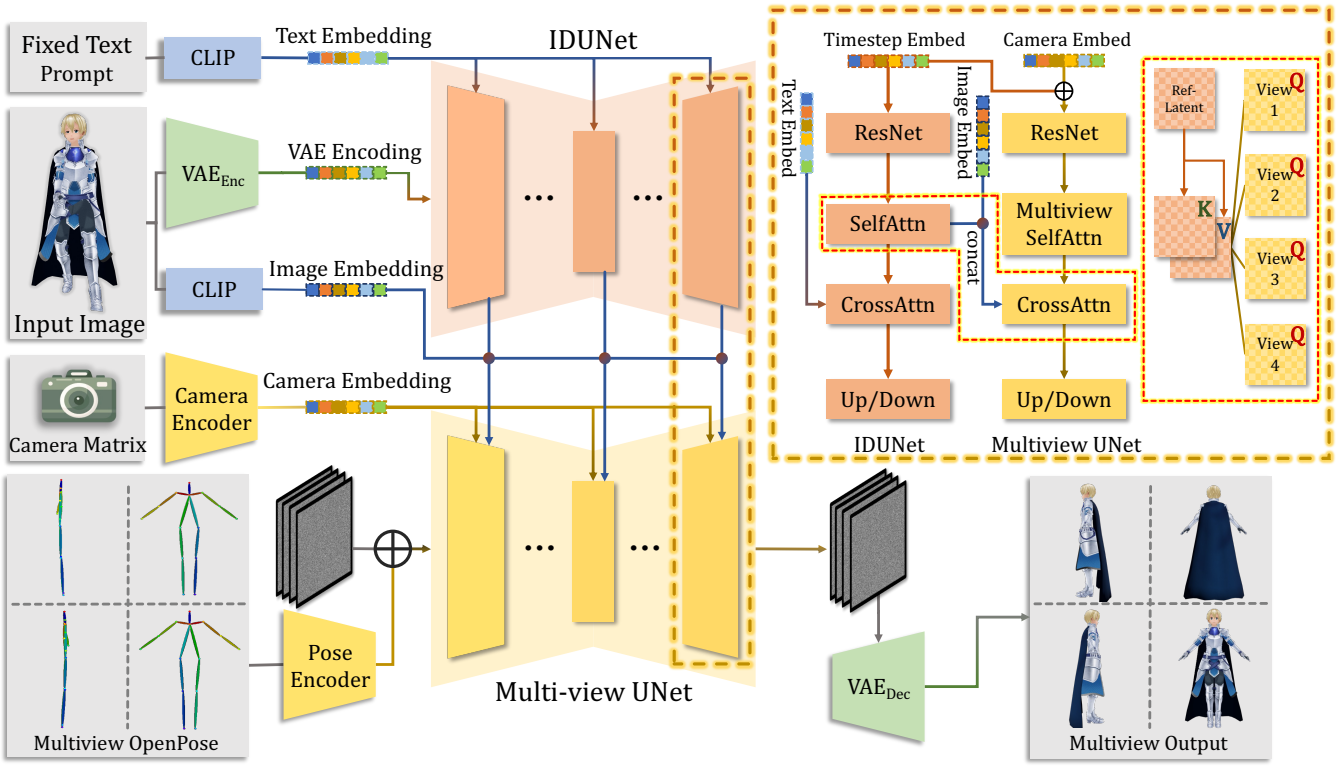

Fig. 3. Pipeline for generating four views of consistent images, showing how our IDUNet extracts local pixel-level features to strengthen the multi-view UNet. Here "Q", "K", and "V" denote the query, key, and value matrix in the attention mechanism.

in the Z-axis and the left and right upper leg to 6° in the Z-axis. All other joint parameters remain untouched. For the *posed* character setting, we download 10 human skeleton animations from Mixamo [MixamoInc. 2009], including sitting, singing, and walking, etc. We randomly select frames from these animations and apply the corresponding motion to the VRM character models. In addition, we also randomize joints of the mouth and eyes to generate a variety of facial expressions, such as winking. We normalize the bounding boxes of the character model to $[-0.5, 0.5]^3$. We configure the camera field of view (FoV) to 40° and set the distance between the camera and the scene origin to 1.5. The character images are rendered with ambient light and directional lighting.

In the training process, we use four A-pose images and a single posed image as a pair because four images in orthographic views already contain sufficient appearance information for a 3D character. Therefore it is natural to render all objects with azimuth angles of $\{0°, 90°, 180°, 270°\}$ and an elevation angle of 0°. To enhance the model's grasp of spatial body layout, we also render three additional groups with random initial azimuth, as depicted in Fig. 2. We also render 4 additional views with completely random azimuth and elevation to fine-tune the generalizable reconstruction model (see Sec. 3.3).

3.2 Multi-view Image Generation and Pose Canonicalization

We now consider how to generate highly consistent multi-view images from the given character image. The overall framework is shown in Fig. 3. We use our IDUNet to transfer patch-level appearance features from the input image to the multi-view denoising UNet. We also introduce a pose embedding network to provide more character layout information to assist the pose canonicalization task.

3.2.1 IDUNet. Our IDUNet aims to retain sufficient features from the original posed image and ensure high consistency between the four generated views. Previous work, IP-Adapter [Ye et al. 2023], adds adapter modules into diffusion UNet structure. Appearance information in input images is transferred to the generated images via the cross-attention mechanism between the input image features and latent features. However, in practice, we observe that IP-Adapter cannot capture the fully detailed texture from input images, as such methods only utilize the global CLIP embeddings of the condition image, which loses pixel-level details during the image encoding, leading to inconsistent results.

To better incorporate features of the condition image, we propose IDUNet, to introduce pixel-level guidance in the generation process. Inspired by ControlNet [Zhang and Agrawala 2023], the structure of

IDUNet is identical to the multi-view UNet. IDUNet takes a fixed text-prompt "best quality" to provide general guidance in the generation process. Unlike ControlNet, to ensure local patch-level interaction between all patches in both the denoised image and the condition image, we leverage the cross-attention between the latent tokens and condition image tokens rather than merely adding them together.

Note that the IDUNet is used to provide pixel-level features into the multi-view UNet, and adding noise to the input condition image severely diminishes the texture detail of the 3D characters. In contrast, a traditional denoising UNet is applied to a noisy image to predict noise according to the timesteps. Thus, we directly adopt a VAE to encode the noise-free input image.

3.2.2 Multi-view UNet. The target of multi-view UNet is to generate multi-view A-pose images with highly consistent appearance from a single posed input image. Within the multi-view UNet, we simultaneously apply the denoising process on the four-view noisy latent $x_{4v} \in \mathrm{R}^{B\times 4\times N\times D}$. Here B, N, and D denote latent batch size, token numbers, and token feature dimension respectively. The multi-view UNet takes the extrinsic camera matrices of the four views as spatial guidance. During the inference stage, the camera poses are set with the fixed azimuths of $\{0°, 90°, 180°, 270°\}$ and the elevation of $0°$. The transformer block in our multi-view UNet consists of a spatial self-attention module and a cross-attention module. As in MVDream [Shi et al. 2023b], the spatial self-attention module directly takes tokens of all four noisy latent x_{4v} and corresponding camera view embeddings. In the spatial self-attention layer, x_{4v} is reshaped into $(B, 4N, D)$ for patch-level cross-view interaction. This design allows the denoising UNet to capture the global relationships across different views, ensuring image generation with high consistency.

In each following cross-attention module, the condition features from IDUNet f_{ID} are concatenated with the CLIP-encoded image features f_{CLIP} to generate final condition features f_{Cond}. Then, a cross-attention layer is used to introduce patch-level interactions into x_{4v}. The whole process is given in Eq. 1 and is illustrated in Fig. 3(above right).

$$f_{\mathrm{Cond}} = \mathrm{concat}(f_{\mathrm{ID}}, f_{\mathrm{CLIP}}) \tag{1}$$

$$x_{4v} = \mathrm{Cross\_Attn}(x_{4v}, f_{\mathrm{Cond}}) \tag{2}$$

Earlier work [Lin et al. 2023b] on the diffusion process shows that applying zero sample-noise-ratio (SNR) at the last timestamp T in training improves the final generation quality because the input noise in the inference stage is pure Gaussian noise. To achieve zero-SNR in the training stage, we manually set SNR_T to zero and linearly scale other Guassian distribution parmaeters β_s. We set the UNet to directly output velocity $v_{\mathrm{pred}} \in \mathrm{R}^{(B,4,N,D)}$ of the multi-view noisy latent and convert it to the noise ϵ_{pred}. The final optimization target is given in Eq. 3.

$$L_{4v} = ||\epsilon_{4v} - \epsilon_{\mathrm{pred}}||_2^2. \tag{3}$$

where ϵ_{4v} is the noise added to the multi-view A-pose images in the diffusion forward pass.

3.2.3 Pose Canonicalization. Combined with IDUNet, our Multi-view diffusion model can successfully generate highly consistent orthographic view images while maintaining detailed features from the prompt image. To enable the diffusion model to achieve character pose canonicalization during the generation process, we jointly train the two UNets with the image pairs from our Anime3D dataset. However, simply training the diffusion network without extra pose constraints will lead to character layout misplacement and the emergence of unrelated body parts. To tackle these problems, we introduce character layouts to the diffusion models for generating A-pose character images by adopting OpenPose [Cao et al. 2021] to predict the pose embedding as an additional condition. The generated embeddings are directly added to the latent noise to aid the diffusion model in learning the relationships between character joints and the generated character layout. Since characters have various body shapes, we leverage three different sets of OpenPose images from our Anime3D dataset and select the one with the highest CLIP score as the input pose condition in the inference stage.

3.3 3D Character Generation

We now describe how we efficiently generate 3D characters from the four-view images generated by our multi-view pose canonicalization diffusion model. As shown in Fig. 4, we adopt a coarse-to-fine process for the 3D character generation task. We first utilize a two-stage transformer-based network to reconstruct the geometry and coarse appearance of the character following the design of LRM [Hong et al. 2023b]. Subsequently, we employ a texture back-projection strategy to quickly improve the texture quality using the generated high-resolution four-view images. Finally, we utilize Poisson Blending [Pérez et al. 2003] to reduce seams on the texture map.

3.3.1 Character Reconstruction with Coarse Texture. Inspired by LRM [Hong et al. 2023b], we utilize a deep transformer network to efficiently reconstruct characters from the four-view images generated by the multi-view diffusion model. While LRM is trained using the Objaverse dataset [Deitke et al. 2023] to allow versatile 3D object generation, it does not sufficiently capture the intricacies of human character layouts. To retain the reconstruction network's ability to process both general 3D objects and stylized characters, we initially pre-train our transformer network on the Objaverse dataset. Then we fine-tune the model with our Anime3D dataset to introduce more priors of human body structure.

Original LRM proposes to mainly train with NeRF [Mildenhall et al. 2022] representation. However, directly extracting geometry from NeRF models often yields noisy surface geometry, which can be problematic for the subsequent use of character meshes in downstream graphics pipelines. Instead, we utilize a two-stage fine-tuning strategy for our reconstruction network. The first stage involves using a triplane NeRF representation similar to LRM, to establish the character's coarse geometry and appearance. In the second stage, we modify the decoder module of our reconstruction network to predict signed distance functions (SDFs) rather than density fields, which allows CharacterGen to achieve smoother and more precise surface geometry.

In addition to MSE loss, we also incorporate mask loss and LPIPS loss [Zhang et al. 2018] to supervise reconstruction appearance. We adopt the binary-cross-entropy loss between ground-truth alpha masks and rendered alpha masks as the mask loss to help the reconstruction model distinguish empty space within input four-view

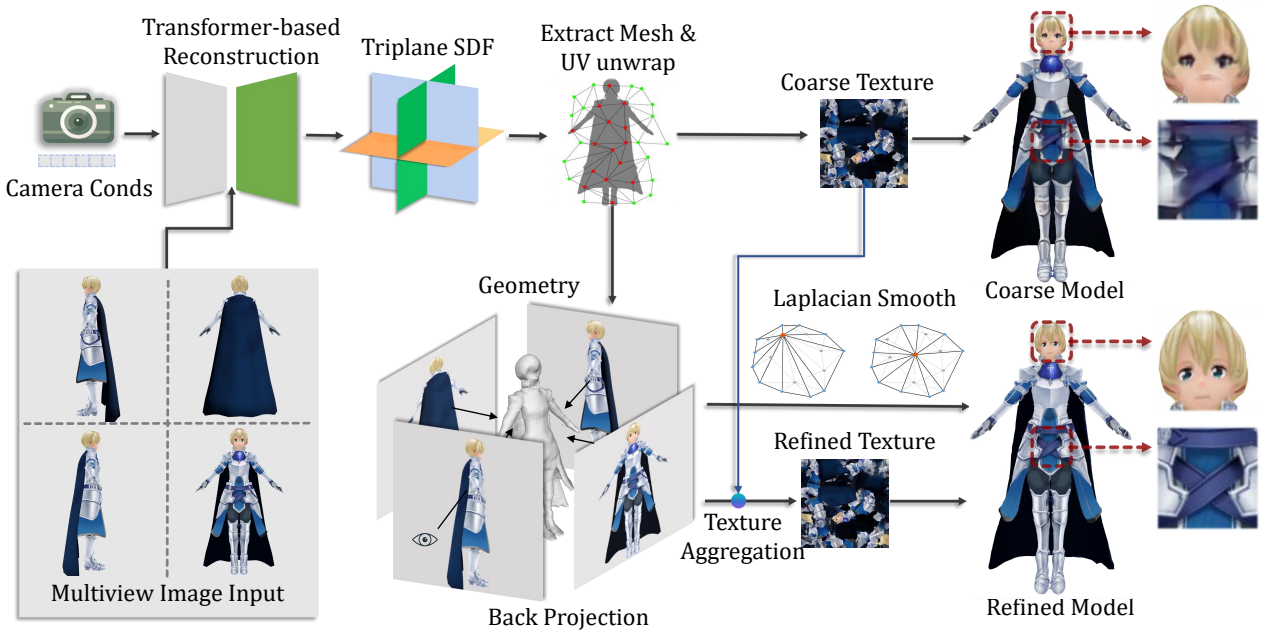

Fig. 4. Pipeline for generating a final refined character mesh from generated multi-view images. In the first stage, we utilize a deep transformer-based network to generate a character with a coarse texture and then use a texture back-projection strategy to enhance the appearance of the generated mesh.

images. LPIPS loss is used to extract perceptual information from input images. The training target is given in Eq. 4.

$$L_{\text{recon}} = \lambda_1 L_{\text{mse}} + \lambda_2 L_{\text{mask}} + \lambda_3 L_{\text{LPIPS}} \tag{4}$$

Here, λ_1, λ_2 and λ_3 are hyperparameters, which are set to 1, 0.1 and 0.5 by default. We leverage Laplacian Smooth on the extracted meshes to further reduce the noise on the surface.

3.3.2 3D Character Refinement. Our reconstruction network can rapidly reconstruct the 3D implicit representation of a character and we can extract the final mesh along with a coarse UV map from the reconstructed tri-plane with DMTet [Shen et al. 2021]. However, the generated mesh still lacks texture details because the DMTet-based extraction process loses appearance information during the UV unwrapping process. To tackle this problem, we further utilize the generated four-view images to improve the quality of the generated texture maps. For efficient rasterization in this step, we employ NvDiffRast [Laine et al. 2020] as the renderer. Since the generated four-view images are in lower resolution than the texture map, multiple texels may be projected onto the same image pixel. During differentiable-rendering-based optimization, the gradients for these texels become noisy. The sparsity of input views compounds the difficulty of fixing these noisy texels, resulting in severe degradation of the output refined texture map. To circumvent this problem, we project the four-view images into texels in texture space and employ a depth test to remove occluded texels. We also notice that directly back-projecting the four-view images onto the mesh leads to noisy texels at the character's body silhouette. We compute the inner products of the four orthogonal camera view directions with the normal texture map. Texels with inner products greater than -0.2 are disregarded to eliminate noise around the silhouette. For texels overlapping in multiple views, we select the back-projected texels with RGB values closest to the coarse texture. Then we utilize Poisson Blending [Pérez et al. 2003] to aggregate projected texels and origin texels to reduce seams in the final textures.

4 EXPERIMENTS

4.1 Implementation Details

We divide our Anime3D dataset into a training set and a testing set in a 50:1 ratio. We use the Stable Diffusion 2.1 model as the base model of both our IDUNet and multi-view UNet. The training process is carried out on 8 NVIDIA A800 GPUs for 3 days on images with 512×512 resolution and another 2 days on images with 768×512 resolution. In each training step, we jointly train both IDUNet and multi-view UNet. We begin by sampling a group of four-view images and a single-condition posed image from Anime3D. Given our intention to generate four images with azimuths of $\{0°, 90°, 180°, 270°\}$, we sample this group with a probability of 0.8. We also include other four-view image groups in our training steps to strengthen the diffusion model's spatial understanding of 3D characters. We sample the front-view posed character images whose azimuth ranges from $[-90°, 90°]$. During the training phase, the posed image is fed into the IDUNet, and the four-view images are sent to the multi-view UNet.

To fine-tune the transformer-based reconstruction model, we first fine-tune for 50 epochs with the NeRF representation and for another 30 epochs on SDF. The fine-tuning process takes 1 day on 8 A800 GPUs. As CharacterGen does not require training in the inference steps, the entire generation pipeline (including four

Table 1. We show the quantitative metrics of both 2D multi-view generation and 3D character generation methods on the test split of Anime3D to evaluate the effectiveness of our CharacterGen.

Methods	SSIM↑	LPIPS↓	FID↓	CD↓
CharacterGen(2D)	**0.901**	**0.086**	**0.019**	-
Zero123	0.768	0.224	1.42	-
Zero123(fine-tuned)	0.813	0.175	1.34	-
SyncDreamer	0.807	0.194	0.396	-
SyncDreamer(fine-tuned)	0.822	0.17	0.37	-
IP-Adapter+SDXL	0.845	0.143	0.074	-
CharacterGen(3D)	**0.898**	**0.093**	**0.032**	**0.001**
Magic123	0.873	0.134	0.116	0.0034
ImageDream	0.886	0.11	0.345	0.002

canonical-pose image generation, 3D character mesh reconstruction, and texture map refinement) can run on a single GPU.

4.2 Results and Comparison

We conduct experiments on both 2D multi-view character image generation and 3D character mesh generation to evaluate the efficiency and effectiveness of our CharacterGen.

4.2.1 2D Multi-view Generation. We test our models on images from the testing split of Anime3D as well as online sources and compare our results to those from Zero123 [Liu et al. 2023c] and SyncDreamer [Liu et al. 2023a]. The comparison results are shown in Fig. 5. It can be seen that, given some difficult body poses, Zero123 and SyncDreamer struggle to preserve enough geometry and appearance information of generated images. Our CharacterGen adeptly performs canonical pose calibration and generates consistent character images across four views, which significantly enhances the subsequent character mesh reconstruction process.

We also conduct experiments on all character images from the test split of Anime3D. We fine-tune the baseline methods on Anime3D for 100 epochs and show the quality metrics in the upper part of Tab. 1. Our calibrated A-pose images are benchmarked against ground-truth A-pose images, whereas images generated by other methods are compared with corresponding posed images. The results evaluate CharacterGen's superior generation quality and its consistency with the multi-view diffusion model.

4.2.2 3D Character Generation. In this section, we compare our generated 3D characters with image-prompt 3D character generation methods. ImageDream [Wang and Shi 2023] and Magic123 [Qian et al. 2023] all utilize the SDS-based optimization. TeCH [Huang et al. 2023b] extends ECON [Xiu et al. 2023] and utilizes DreamBooth [Ruiz et al. 2023] to achieve image-prompt generation. We visualize the results in Fig. 6. We also conduct quantitative experiments on the 3D generation results on the test split of Anime3D. The texture quality metrics are obtained by comparing rendered images and ground-truth images across the four orthogonal views. We choose Chamfer Distance(CD) as the metric to evaluate the geometry quality of generated meshes. We normalize the meshes to $[-0.5, 0.5]^3$ for calculating CD. The quantitative results are listed in the lower part of Tab. 1.

Table 2. Time to generate a single 3D character. Models loading time is excluded for all methods.

Methods	Time
CharacterGen	1min
Magic123 [Qian et al. 2023]	70min
ImageDream [Wang and Shi 2023]	45min
TeCH [Huang et al. 2023b]	270min

It can be observed that our CharacterGen effectively avoids "Janus" problem thanks to our robust four-view reconstruction mechanism. Our generated 3D character meshes also exhibit satisfactory appearance for unseen body parts, with resourceful back-view and side-view priors from our Anime3D. Most 3D characters generated by other methods suffer from several mesh face cohesion problems, which makes it extremely hard to rig and animate these characters. CharacterGen can successfully generate canonical pose meshes from characters with tricky poses, which facilitates downstream graphic applications. We also evaluate other methods using A-pose character images canonicalized by Animate Anyone [Hu et al. 2023]. Please refer to the supplementary materials for further details.

4.2.3 Comparison with IP-Adapter. Previous work IP-Adapter [Ye et al. 2023] also supports the image-prompt generation task by incorporating adapter modules into diffusion models. We notice that IP-Adapter does not include pre-trained models for Stable Diffusion 2.1, which is the base model of our multi-view UNet. Alternatively, we train an SDXL-base model to generate 2×2 grid character images in four views with the azimuth of $\{0°, 90°, 180°, 270°\}$ for 100 epochs. The grid images are organized with a resolution of 1024×1024, which is the standard resolution of SDXL. Then we integrate official pre-trained IP-Adapter-SDXL into the base SDXL model to obtain image-conditioned multi-view results. We show the visualization results in Fig. 7 and quantitative results in Tab. 1. It can be observed that IP-Adapter cannot preserve detailed appearance information and may fail to generate correct character layout, while CharacterGen can effectively generate high-consistent multi-view character images with our IDUNet.

4.2.4 Generation Speed. We compare the time needed to generate a single 3D character mesh by our method and other image-prompted 3D generation methods, with the comparison results detailed in Tab. 2. Our method is significantly faster than other alternatives. SyncDreamer and Zero123 are used to generate multi-view images. Their time cost varies based on the chosen 3D reconstruction methods. The default NeuS [Wang et al. 2021a] reconstruction takes about 10 minutes.

4.3 User Study

To better evaluate the robust generation ability of our CharacterGen, we collect 15 generated four-view character images and 10 character meshes for our user study. We compare results generated by CharacterGen with other methods mentioned in Sec. 4.2. We

Fig. 5. We compare our generated four A-pose character images with other methods. The azimuths for all examples are set as $\{0°, 90°, 180°, 270°\}$. © kinoko7

Fig. 6. We compare the appearance and geometry of our generated 3D characters with other methods. © kinoko7

Table 3. Statistics of the user study. We display the voting results for both 2D multi-view images and 3D textured character meshes.

metric	CharacterGen	Zero123(2D)	SyncDreamer(2D)	Magic123(3D)	ImageDream(3D)
2D multi-view style consistency	**85.4%**	10.5%	4.1%	-	-
2D multi-view consistency	**81.0%**	17.1%	1.9%	-	-
3D character geometry quality	**78.6%**	-	-	2.86%	18.6%
3D character texture quality	**87.1%**	-	-	1.9%	11.0%

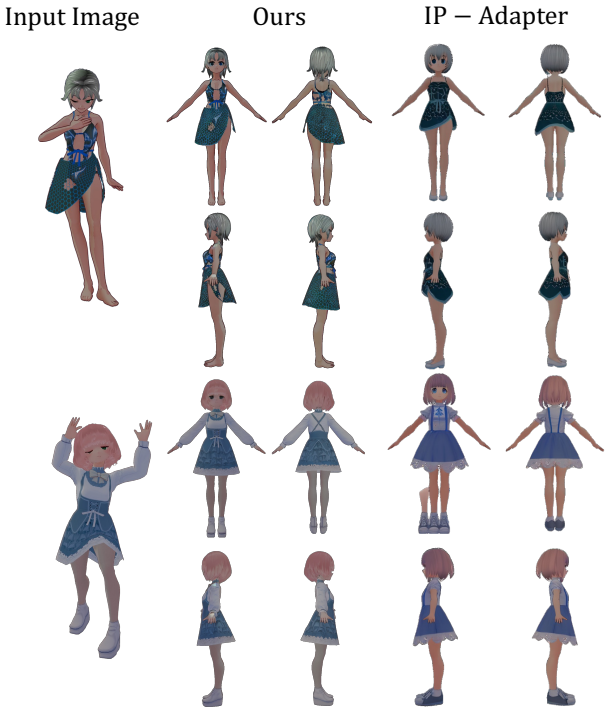

Fig. 7. We compare the results of CharacterGen and IP-Adapter-SDXL.

Table 4. We utilize CLIP score to assess style similarity between generated characters and given condition images with 2D multi-view images and 3D rendered images.

method(2D)	CLIP score	method(3D)	CLIP score
CharacterGen	**83.69**	CharacterGen	**79.77**
Zero123	79.02	Magic123	74.71
SyncDreamer	75.41	ImageDream	73.65

ask 21 volunteers to first evaluate the style consistency between the condition images and the generated four-view images, as well as the spatial consistency within the generated multi-view images. Then, they are required to assess the geometry quality and texture quality of generated 3D character models. For each example, the volunteers are asked to vote for the one they consider to be the best. As shown in Tab 3, our CharacterGen receives a significantly higher preference compared to other methods for both 2D and 3D generation tasks.

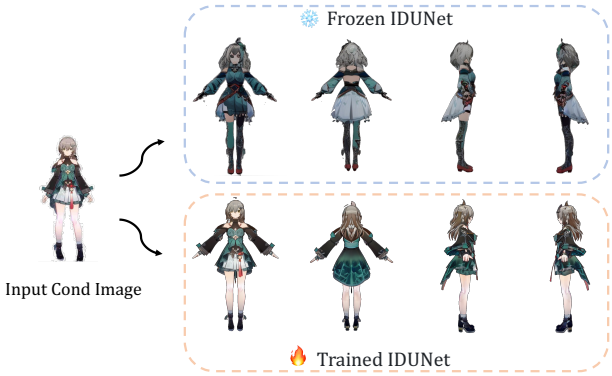

Fig. 8. Frozen IDUNet cannot extract sufficient appearance information from the prompt image and generates dissimilar images.

We additionally quantitatively evaluate the coherence between our generated results and the input images the same as in the user study. We adopt the CLIP score as the metric and utilize ViT-B/32 as the backbone model. For 2D multi-view image generation, we calculate the CLIP scores of the input image and four generated images. For 3D character generation, we first render the 3D representation at an azimuth interval of 3°, and then compute the CLIP scores between the input image and all the rendered frames. As displayed in Tab. 4, our CharacterGen gains superior results in both 2D and 3D generation, evaluating the robust appearance modeling ability of our IDUNet, multi-view UNet, and the subsequent reconstruction model.

4.4 Ablation Study

4.4.1 IDUNet. To demonstrate the importance of jointly training the IDUNet, we train CharacterGen network while freezing IDUNet with the pre-trained stable diffusion 2.1 model. The generated four-view images are shown in Fig. 8. The results reveal that the generated images fail to preserve sufficient features from the input images, resulting in reduced similarity. This shows the necessity of jointly fine-tuning the IDUNet with clean posed images to enhance its ability to extract detailed clothes and facial appearance.

4.4.2 Pose Embedding Network. The pose embedding network plays a crucial role in keeping character layouts in the generated four-view images. We generate additional sets of images without the pose embedding network and display representative results in Fig. 9. It can be observed that the generated character images may not be located in the middle of the image in the absence of the pose embedding network. Furthermore, the lack of layout guidance can

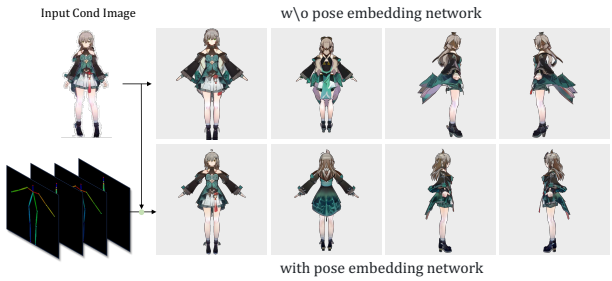

Fig. 9. Without the pose embedding network, the generated characters may be misplaced.

lead to the generation of inconsistent clothing parts, which could compromise 3D reconstruction in the subsequent step.

4.5 Applications

CharacterGen can generate A-pose 3D characters with detailed texture maps, thereby simplifying the subsequent rigging process. We employ AccuRig [actorcore 2023] to automatically rig our generated character meshes. The rigged 3D characters can be readily utilized as animated 3D assets in various domains. We render various animated rigged models in Warudo [HakuyaLabs 2023] and present some results in Fig. 10.

Fig. 10. We rig the generated characters and utilize them as 3D assets in downstream applications.

To better evaluate how using an A-pose character aids skeleton rigging and animation process, we rig two 3D character meshes generated by our CharacterGen and ImageDream [Wang and Shi 2023] and visualize the animated characters in Fig. 11. It can be observed that non-A-pose characters encounter severe mesh cohesion problem and the body structure is severely distorted while our A-posed character can be successfully animated.

5 LIMITATIONS AND DISCUSSION

While our method can generate 3D characters from a single input image in an arbitrary pose, certain limitations still exist. For the four-view A-pose image generation step, our method may not retain enough information when the character is in an extreme pose or is rendered from a non-common viewpoint.

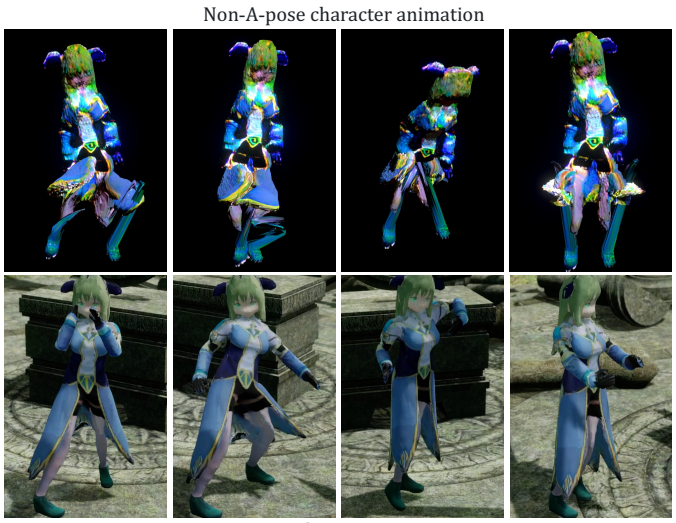

Fig. 11. We compare CharacterGen's animated 3D characters with ImageDream [Wang and Shi 2023]

As for future works, the integration of additional non-photorealistic rendering (NPR) techniques into the texture refinement stage may further enhance the texture quality of the generated characters. Moreover, leveraging our trained multi-view UNet structure, it may be possible to incorporate the SDS optimization method to achieve 3D character generation with superior geometry quality.

6 CONCLUSIONS

This paper proposes CharacterGen, a novel and efficient image-prompt 3D character generation framework. We compile a new multi-pose, stylized character dataset Anime3D to train our pipeline. Our designs include IDUNet, which extracts patch-level features from the input condition image to guide multi-view A-pose character image generation. Subsequently, we utilize a transformer-based network to reconstruct 3D character meshes and propose to utilize the texture back-projection refinement strategy to further improve the appearance of the reconstructed character meshes. Experiments demonstrate that CharacterGen can efficiently generate high-quality 3D characters suitable for multiple downstream applications.

ACKNOWLEDGMENTS

This work is supported by the National Science and Technology Major Project (2021ZD0112902), the National Natural Science Foundation of China (Grant No. 62220106003), and Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology. The authors would like to thank Ralph R. Martin, Yuan-Chen Guo, and Yang-Guang Li for helpful discussion.

REFERENCES

actorcore. 2023. *accurig, a software for automatic character rigging.* https://actorcore.reallusion.com/auto-rig/accurig

Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. 2021. imGHUM: Implicit Generative Models of 3D Human Shape and Articulated Pose. In *2021 IEEE/CVF*

International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, 5441–5450. https://doi.org/10.1109/ICCV48922.2021.00541

Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. 2022. Pretrain, Self-train, Distill: A simple recipe for Supersizing 3D Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 3763–3772. https://doi.org/10.1109/CVPR52688.2022.00375

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. 2023. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. *CoRR* abs/2304.00916 (2023). https://doi.org/10.48550/ARXIV.2304.00916 arXiv:2304.00916

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 1 (2021), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

Sihun Cha, Kwanggyoon Seo, Amirsaman Ashtari, and Junyong Noh. 2023. Generating Texture for 3D Human Avatar from a Single Image using Sampling and Refinement Networks. *Comput. Graph. Forum* 42, 2 (2023), 385–396. https://doi.org/10.1111/CGF.14769

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023a. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. *CoRR* abs/2303.13873 (2023). https://doi.org/10.48550/ARXIV.2303.13873 arXiv:2303.13873

Shuhong Chen, Kevin Zhang, Yichun Shi, Heng Wang, Yiheng Zhu, Guoxian Song, Sizhe An, Janus Kristjansson, Xiao Yang, and Matthias Zwicker. 2023b. PAniC-3D: Stylized Single-view 3D Reconstruction from Portraits of Anime Characters. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 21068–21077. https://doi.org/10.1109/CVPR52729.2023.02018

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A Universe of Annotated 3D Objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 13142–13153. https://doi.org/10.1109/CVPR52729.2023.01263

Congyue Deng, Chiyu Max Jiang, Charles R. Qi, Xinchen Yan, Yin Zhou, Leonidas J. Guibas, and Dragomir Anguelov. 2023. NeRDi: Single-View NeRF Synthesis with Language-Guided Diffusion as General Image Priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 20637–20647. https://doi.org/10.1109/CVPR52729.2023.01977

Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. 2023. HyperDiffusion: Generating Implicit Neural Fields with Weight-Space Diffusion. *CoRR* abs/2303.17015 (2023). https://doi.org/10.48550/ARXIV.2303.17015 arXiv:2303.17015

HakuyaLabs. 2023. *warudo, a 3D virtual image live broadcast software.* https://warudo.app/

Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. 2023a. EVA3D: Compositional 3D Human Generation from 2D Image Collections. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=g7U9jD_2CUr

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. *ACM Trans. Graph.* 41, 4 (2022), 161:1–161:19. https://doi.org/10.1145/3528223.3530094

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2023b. LRM: Large Reconstruction Model for Single Image to 3D. *CoRR* abs/2311.04400 (2023). https://doi.org/10.48550/ARXIV.2311.04400 arXiv:2311.04400

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. 2023. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. *CoRR* abs/2311.17117 (2023). https://doi.org/10.48550/ARXIV.2311.17117 arXiv:2311.17117

Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. 2023a. DreamWaltz: Make a Scene with Complex 3D Animatable Avatars. *CoRR* abs/2305.12529 (2023). https://doi.org/10.48550/ARXIV.2305.12529 arXiv:2305.12529

Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. 2023b. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. *CoRR* abs/2308.08545 (2023). https://doi.org/10.48550/ARXIV.2308.08545 arXiv:2308.08545

Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. 2022. Neural Wavelet-domain Diffusion for 3D Shape Generation. In *SIGGRAPH Asia 2022 Conference Papers, SA 2022, Daegu, Republic of Korea, December 6-9, 2022*, Soon Ki Jung, Jehee Lee, and Adam W. Bargteil (Eds.). ACM, 24:1–24:9. https://doi.org/10.1145/3550469.3555394

Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2023. AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control. *CoRR* abs/2303.17606 (2023). https://doi.org/10.48550/ARXIV.2303.17606 arXiv:2303.17606

Heewoo Jun and Alex Nichol. 2023. Shap-E: Generating Conditional 3D Implicit Functions. *CoRR* abs/2305.02463 (2023). https://doi.org/10.48550/ARXIV.2305.02463 arXiv:2305.02463

Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. 2023. DreamHuman: Animatable 3D Avatars from Text. *CoRR* abs/2306.09329 (2023). https://doi.org/10.48550/ARXIV.2306.09329 arXiv:2306.09329

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. 2020. Modular primitives for high-performance differentiable rendering. *ACM Trans. Graph.* 39, 6 (2020), 194:1–194:14. https://doi.org/10.1145/3414685.3417861

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. 2023. Instant3D: Fast Text-to-3D with Sparse-View Generation and Large Reconstruction Model. *CoRR* abs/2311.06214 (2023). https://doi.org/10.48550/ARXIV.2311.06214 arXiv:2311.06214

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813

Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. 2023. TADA! Text to Animatable Digital Avatars. *CoRR* abs/2308.10899 (2023). https://doi.org/10.48550/ARXIV.2308.10899 arXiv:2308.10899

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023a. Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 300–309. https://doi.org/10.1109/CVPR52729.2023.00037

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. 2023b. Common Diffusion Noise Schedules and Sample Steps are Flawed. *CoRR* abs/2305.08891 (2023). https://doi.org/10.48550/ARXIV.2305.08891 arXiv:2305.08891

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. 2023b. One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion. *CoRR* abs/2311.07885 (2023). https://doi.org/10.48550/ARXIV.2311.07885 arXiv:2311.07885

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. 2023d. One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. *CoRR* abs/2306.16928 (2023). https://doi.org/10.48550/ARXIV.2306.16928 arXiv:2306.16928

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023c. Zero-1-to-3: Zero-shot One Image to 3D Object. *CoRR* abs/2303.11328 (2023). https://doi.org/10.48550/ARXIV.2303.11328 arXiv:2303.11328

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023a. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. *CoRR* abs/2309.03453 (2023). https://doi.org/10.48550/ARXIV.2309.03453 arXiv:2309.03453

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. 2023. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. *CoRR* abs/2310.15008 (2023). https://doi.org/10.48550/ARXIV.2310.15008 arXiv:2310.15008

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34, 6 (2015), 248:1–248:16. https://doi.org/10.1145/2816795.2818013

Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. 2023. RealFusion 360° Reconstruction of Any Object from a Single Image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 8446–8455. https://doi.org/10.1109/CVPR52729.2023.00816

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2023. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 12663–12673. https://doi.org/10.1109/CVPR52729.2023.01218

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2022), 99–106. https://doi.org/10.1145/3503250

MixamoInc. 2009. Mixamo's online services. https://www.mixamo.com/

Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. 2023. DiffRF: Rendering-Guided 3D Radiance Field Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 4328–4338. https://doi.org/10.1109/CVPR52729.2023.00421

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *CoRR* abs/2212.08751 (2022). https://doi.org/10.48550/ARXIV.2212.08751 arXiv:2212.08751

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 10975–10985. https://doi.org/10.1109/CVPR.2019.01123

Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson image editing. *ACM Trans. Graph.* 22, 3 (2003), 313–318. https://doi.org/10.1145/882262.882269

Pixiv. 2019. *VRM tools of three.js.* https://github.com/pixiv/three-vrm

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=FjNys5c7VyY

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. 2023. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. *CoRR* abs/2306.17843 (2023). https://doi.org/10.48550/ARXIV.2306.17843 arXiv:2306.17843

Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan T. Barron, Yuanzhen Li, and Varun Jampani. 2023. DreamBooth3D: Subject-Driven Text-to-3D Generation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2349–2359. https://doi.org/10.1109/ICCV51070.2023.00223

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 22500–22510. https://doi.org/10.1109/CVPR52729.2023.02155

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2304–2314. https://doi.org/10.1109/ICCV.2019.00239

Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 81–90. https://doi.org/10.1109/CVPR42600.2020.00016

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 6087–6101. https://proceedings.neurips.cc/paper/2021/hash/30a237d18c50f563cba4531f1db44acf-Abstract.html

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. 2023a. Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model. *CoRR* abs/2310.15110 (2023). https://doi.org/10.48550/ARXIV.2310.15110 arXiv:2310.15110

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. 2023b. MVDream: Multi-view Diffusion for 3D Generation. *CoRR* abs/2308.16512 (2023). https://doi.org/10.48550/ARXIV.2308.16512 arXiv:2308.16512

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023a. Make-It-3D: High-Fidelity 3D Creation from A Single Image with Diffusion Prior. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 22762–22772. https://doi.org/10.1109/ICCV51070.2023.02086

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. 2023b. MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion. *CoRR* abs/2307.01097 (2023). https://doi.org/10.48550/ARXIV.2307.01097 arXiv:2307.01097

Anh Thai, Stefan Stojanov, Vijay Upadhya, and James M. Rehg. 2021. 3D Reconstruction of Novel Object Shapes from Single Images. In *International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021*. IEEE, 85–95. https://doi.org/10.1109/3DV53792.2021.00019

VRoid. 2022. VRoid Hub. https://vroid.com/

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. 2023a. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 12619–12629. https://doi.org/10.1109/CVPR52729.2023.01214

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Hang Yu, Wei Liu, Xiangyang Xue, and Yu-Gang Jiang. 2021b. Pixel2Mesh: 3D Mesh Model Generation via Image Guided Deformation. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 10 (2021), 3600–3613. https://doi.org/10.1109/TPAMI.2020.2984232

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021a. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 27171–27183. https://proceedings.neurips.cc/paper/2021/hash/e41e164f7485ec4a28741a2d0ea41c74-Abstract.html

Peng Wang and Yichun Shi. 2023. ImageDream: Image-Prompt Multi-view Diffusion for 3D Generation. *CoRR* abs/2312.02201 (2023). https://doi.org/10.48550/ARXIV.2312.02201 arXiv:2312.02201

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. 2023c. RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 4563–4573. https://doi.org/10.1109/CVPR52729.2023.00443

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *CoRR* abs/2305.16213 (2023). https://doi.org/10.48550/ARXIV.2305.16213 arXiv:2305.16213

Chao Wen, Yinda Zhang, Chenjie Cao, Zhuwen Li, Xiangyang Xue, and Yanwei Fu. 2023. Pixel2Mesh++: 3D Mesh Generation and Refinement From Multi-View Images. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 2 (2023), 2166–2180. https://doi.org/10.1109/TPAMI.2022.3169735

Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. 2023a. Multiview Compressive Coding for 3D Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 9065–9075. https://doi.org/10.1109/CVPR52729.2023.00875

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. 2023b. OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 803–814. https://doi.org/10.1109/CVPR52729.2023.00084

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 512–523. https://doi.org/10.1109/CVPR52729.2023.00057

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. 2022. ICON: Implicit Clothed humans Obtained from Normals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 13286–13296. https://doi.org/10.1109/CVPR52688.2022.01294

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *CoRR* abs/2308.06721 (2023). https://doi.org/10.48550/ARXIV.2308.06721 arXiv:2308.06721

Paul Yoo, Jiaxian Guo, Yutaka Matsuo, and Shixiang Shane Gu. 2023. DreamSparse: Escaping from Plato's Cave with 2D Frozen Diffusion Model Given Sparse Views. *CoRR* abs/2306.03414 (2023). https://doi.org/10.48550/ARXIV.2306.03414 arXiv:2306.03414

Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. 2022. LION: Latent Point Diffusion Models for 3D Shape Generation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/40e56dabe12095a5fc44a6e4c3835948-Abstract-Conference.html

Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 2023b. 3DShape2VecSet: A 3D Shape Representation for Neural Fields and Generative Diffusion Models. *ACM Trans. Graph.* 42, 4 (2023), 92:1–92:16. https://doi.org/10.1145/3592442

Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. 2023a. AvatarVerse: High-quality & Stable 3D Avatar Creation from Text and Pose. *CoRR* abs/2308.03610 (2023). https://doi.org/10.48550/ARXIV.2308.03610 arXiv:2308.03610

Jianfeng Zhang, Xuanmeng Zhang, Huichao Zhang, Jun Hao Liew, Chenxu Zhang, Yi Yang, and Jiashi Feng. 2023c. AvatarStudio: High-fidelity and Animatable 3D Avatar Creation from Text. *CoRR* abs/2311.17917 (2023). https://doi.org/10.48550/ARXIV.2311.17917 arXiv:2311.17917

Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *CoRR* abs/2302.05543 (2023). https://doi.org/10.48550/ARXIV.2302.05543 arXiv:2302.05543

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 586–595. https://doi.org/10.1109/CVPR.2018.00068

Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. 2022. PaMIR: Parametric Model-Conditioned Implicit Representation for Image-Based Human Reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 6 (2022), 3170–3184. https://doi.org/10.1109/TPAMI.2021.3050505