DRiVE: Diffusion-based Rigging Empowers Generation of Versatile and Expressive Characters

Mingze Sun[1*] Junhao Chen[1*] Junting Dong[2†] Yurun Chen[1] Xinyu Jiang[1] Shiwei Mao[1] Puhua Jiang[1]
Jingbo Wang[2] Bo Dai[2] Ruqi Huang[1†]
[1]Tsinghua Shenzhen International Graduate School, China
[2]Shanghai AI Laboratory, China

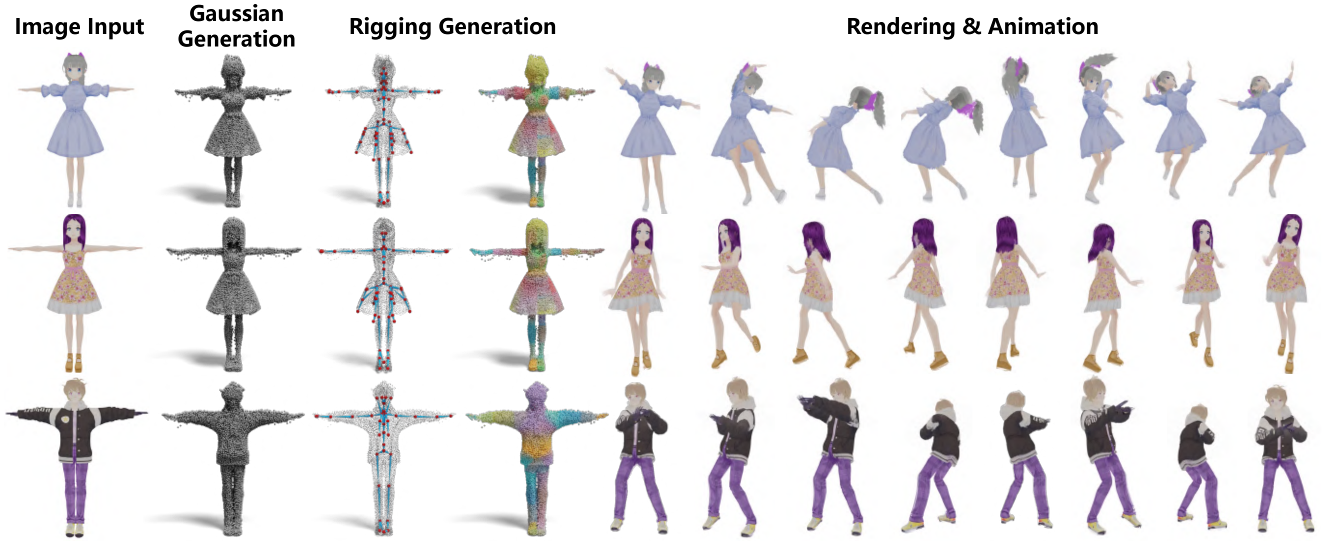

Figure 1. We propose **DRiVE**, a pipeline that generates 3D Gaussian from a single image along with the corresponding skeleton (including hair and clothing) and skinning, enabling precise control over 3D Gaussian to render high-quality, controllable, and 3D consistent videos.

Abstract

Recent advances in generative models have enabled high-quality 3D character reconstruction from multi-modal. However, animating these generated characters remains a challenging task, especially for complex elements like garments and hair, due to the lack of large-scale datasets and effective rigging methods. To address this gap, we curate `AnimeRig`*, a large-scale dataset with detailed skeleton and skinning annotations. Building upon this, we propose **DRiVE**, a novel framework for generating and rigging 3D human characters with intricate structures. Unlike existing methods, DRiVE utilizes a 3D Gaussian representation, facilitating efficient animation and high-quality rendering. We further introduce GSDiff, a 3D Gaussian-based diffusion module that predicts joint positions as spatial distributions, overcoming the limitations of regression-based approaches. Extensive experiments demonstrate that DRiVE achieves precise rigging results, enabling realistic dynamics for clothing and hair, and surpassing previous methods in both quality and versatility. The code and dataset will be made public for academic use at https://DRiVEAvatar.github.io/.*

1. Introduction

Crafting and animating 3D human characters has long been a critical task in an array of applications, including film and video game making, AR/VR, and human-centric robotics, to name a few. Notably, the rapid development of generative models opens numerous new opportunities for research on this classic task. In contrast to the traditional manual and time-consuming crafting pipeline, nowadays one can *create* high-quality graphical human models with *rich structures* (*e.g.,* intricate garment and hairs) from a single image [31, 40, 41, 53, 54], a piece of text prompt [15], or a video clip [13, 55] automatically and efficiently.

[*] Indicates Equal Contribution. [†] Indicates Corresponding Author.

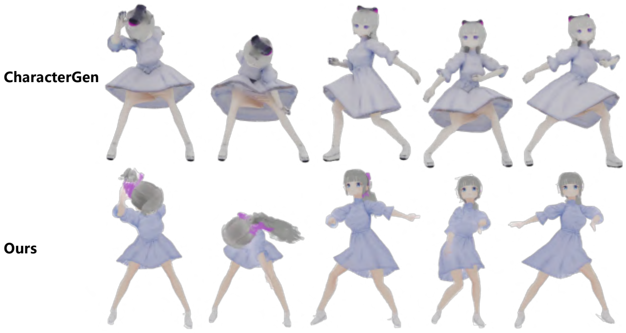

Figure 2. We compare our method with CharacterGen [31] animation results. Since **DRiVE** explicitly models clothing and hair, it generates more natural and realistic animations. Additionally, 3D Gaussian-based rendering achieves higher quality than mesh.

Nevertheless, the development of animating created 3D characters is relatively lagged – the current main focus is still positioned on addressing motions of the human body, which is typically based on 2D [5] or 3D parametric human models [27, 30]. The lack of tailored-for designs for external parts beyond the human body significantly limits the downstream applications. For instance, as shown in Fig. 2, naively animating a female character in a dress and wearing her hair in a ponytail with a body skeleton would lead to unrealistic rendering since her dress and hair are rigidly stuck to the body.

In light of the above challenge, we present **DRiVE**, to the best of our knowledge, the first framework towards *generation and animation* of 3D human characters with rich structures beyond body parts. We first discuss the critical design choices in building DRiVE as follows.

Animation Representation: While it is tempting to shoot for an automatic framework in a data-driven way, we take a step back and resort to computer graphics for a balance between automation and control degree. More specifically, we aim for *rigging* the generated characters with a set of skeletons including rich structures such as hair and clothing and the corresponding skinning. It is worth noting that SMPL and the follow-ups are essential rigging methods as well. The key difference in-between is that our rigging target is of a heterogenous structure, which is far more challenging than the former.

Generation representation: There is an amount of 3D shape representations of choice for generative models. As we desire efficient animation and rendering, we rule out point clouds and implicit representations despite their advantage in representing the complex topology of garments. It may then seem natural to settle down at polygonal mesh, due to its popularity in the prior academic and industrial efforts on rigging. However, we identify 3D Gaussian [18] as a better alternative for the following reasons: 1) making use of mesh-based rigging frameworks requires meshing quality (*e.g.,* watertight [2]), which is hard to guarantee in generation; 2) even geometry of mesh were given a prior, producing high-quality texture maps remains an open problem [66] (also see in Fig. 5).

To conclude, DRiVE generates high-quality 3D Gaussian from low-level input such as a single image or text prompt. Then it assigns rigging information automatically to the 3D Gaussians, which can be further used for fine-grained control over the body but also external structures.

Two core challenges surface in our design. First, to our knowledge, there does not exist any large-scale dataset for rigging characters with rich structures in general, letting alone being specific to 3D Gaussian; Second, apart from the essential difficulty of learning heterogeneous skeletons from geometries, our generated 3D Gaussian lacks geometric structure, which has been considered important in estimating skeletons [58, 59]. In response to **Data Insufficiency**, we curate `AnimeRig`, a large set of 9420 meshes with calibrated rigging annotations (including skeleton and skinning). Then, to adapt to our representation, we fine-tune LGM [46] to generate 3D Gaussian from a single image. Finally, we perform label transfer from the original meshes to the corresponding 3D Gaussian. We remark that the fine-tuned LGM also serves as our final generation model; Regarding **Difficulty of Learning Skeletons on 3D Gaussian**, we propose a novel diffusion module, GSDiff, to de-noise for *joint positions* conditioned on the input 3D Gaussian. In contrast to the previous regression-based methods, GSDiff treats joint positions as spatial distributions, taming the learning difficulty. Additionally, for fully exploiting 3D Gaussian, we separately extract geometric and appearance features from the means and canonical multi-view renderings for diffusion. The accurate joint prediction then lays a good basis for further estimating bone connection and skinning weight, leading to high-quality rigging results.

Last but not least, beyond our best expectations, the above framework can be easily extended to rigging (potentially textured) meshes, which on its own is a tough task. More specifically, we *drop* the edge connection of the meshes, and train GSDiff using the vertex sets and canonical multi-view renderings as we train on 3D Gaussian.

We summarize our main contributions as follows:

1. We introduce a pipeline that generates rigged 3D models from multimodal inputs, enabling the detailed modeling of complex elements such as hair and clothing, thereby producing high-quality, free-viewpoint rendered videos.
2. We propose GSDiff, a novel 3D Gaussian-based diffusion network, to accurately predict joint positions. For 3D Gaussian conditional inputs, we specifically design a tailored conditioning approach.
3. By fully exploiting 3D Gaussian geometric and appearance information, we enhance the rigging process, enabling precise skeleton binding and skinning estimation.

4. We introduce AnimeRig, a large-scale character rigging dataset, and achieve state-of-the-art results in the skeleton and skinning predictions on this dataset. Our results can be directly integrated into animation pipelines, significantly reducing the complexity of animators' workflows.

2. Related Works

2.1. 3D Avatar Rigging

In 3D avatar rigging, the task traditionally requires manual rigging by designers, which is both time-consuming and tedious. In recent years, with the advancement of machine learning technologies, automatic rigging methods have begun to emerge. Notably, Neural body [33] first proposes using the SMPL to generate dynamic 3d human models automatically. This groundbreaking research laid the foundation for animatable 3d human generation using the SMPL model for [4, 14, 44, 60]. However, SMPL model focuses entirely on the human body, which cannot effectively drive clothing, hair, and other elements, resulting in unrealistic dynamic effects. To this end, [37] and [64] do not directly use SMPL but instead learn from rigging data with labeled human bodies. However, their labeled skeleton data still has a similar topology to SMPL, leading to similar issues during animation as seen in SMPL-based methods.

To address the aforementioned issues, we focus on generating heterogeneous skeletons. Current mainstream methods can be divided into two categories: optimization-based methods and learning-based methods. CASA [52] is the first to propose jointly inferring articulated skeletal shapes and rigging through optimization. Subsequent follow-ups combine techniques such as dynamic NeRF [45, 61, 62] and dual-phase optimization [67] to improve the quality of 3D object reconstruction and rigging. However, optimization-based methods can only be applied on a per-case basis, lacking generalization capabilities and therefore are costly for large-scale data processing.

Recent works have utilized learning-based methods to generate heterogeneous skeletons [2, 28, 58, 59, 63] and skinning [9, 21]. Among them, RigNet [59] takes a mesh as input and designs networks to predict the positions of joints, bone connections, and skinning separately. For joint estimation, RigNet first predicts offsets using a regression-based approach and then performs clustering in a differentiable manner, determining joint positions based on the cluster centers. This sequential pipeline is relatively complex and prone to error accumulation. Moreover, predictions based on regression methods tend to lack generalization capability. We propose a novel pipeline for joint prediction based on conditional diffusion. By learning the distribution of joints through diffusion, our method enables the generation of complex heterogeneous skeletons.

2.2. 3D Avatar Generation

The development of general 3D generation technologies has significantly enhanced the realism and detail of avatars created from various inputs, ranging from objects to human figures [3, 6, 12, 19, 20, 22, 23, 25, 26, 43, 47, 50, 56, 69]. However, human-centered 3D reconstruction methods focus on high-fidelity digitization and reconstruction of clothed humans from minimal inputs [7, 10, 11, 14, 16, 17, 32, 40, 41, 53–55, 70]. Recent approaches have taken single-image 3D reconstruction of anime characters to new heights using cartoon datasets [31]. The Gaussian-based approach offers better rendering quality compared to the mesh-based approach [13, 24, 42, 57]. Despite these advancements, existing methods often overlook skeletal integration and typically use SMPL (Skinned Multi-Person Linear Model) [27, 30] as the driving skeleton, which can cause unnatural behavior in elements like skirts. To address this issue, we propose a method incorporating heterogeneous skeletons into Gaussian representations, allowing for a more realistic simulation of clothing and hair movements in sync with the avatar's body, which is crucial for gaming applications.

3. Dataset Construction

We construct our AnimeRig dataset with the following main stages and defer the technical details to Supp. Mat.:

1. **Data collection:** We first curate a large set of 13746 textured meshes with initial rigging annotations from VRoidHub [1], then we filter out the non-humanoid ones and ask artists to manually repair the significant errors, resulting in a subset of 9420 meshes with reliable annotations. Moreover, each joint is accompanied by some semantic label, such as "J_Bip_L_Hand, J_Sec_R_SkirtBack1".
2. **Conversion to 3D Gaussian:** Then we convert the rigged meshes to 3D Gaussian representation. Specifically, we resort to fine-tuning LGM [46] with images rendered from the given meshes, which leads to a conversion route as input mesh → 4-view image rendering of the mesh → 3D Gaussian generated by fine-tuned LGM.
3. **Label Transfer to 3D Gaussian:** Since the faithfully generated 3D Gaussian do not necessarily align with the mesh input, we further employ a scaled Iterative Closest Point (ICP) algorithm [8] to register the mesh to the corresponding 3D Gaussian, with naturally transform annotations from the source meshes to the 3D Gaussian. An example is shown in Fig. 3(b).

4. Methodology

In this section, we demonstrate the overall pipeline of our rigging system, which, trained on our AnimeRig dataset,

[1] https://hub.vroid.com/en

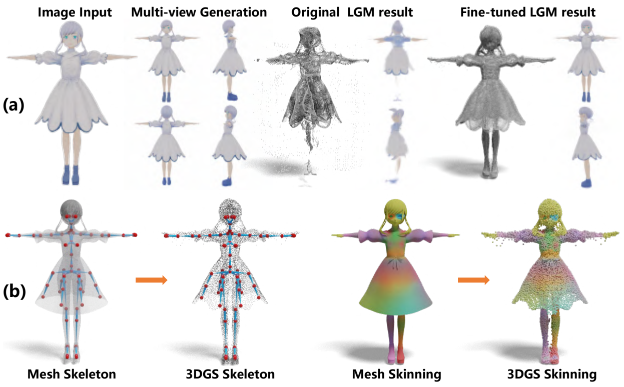

Figure 3. We show the comparison results of LGM before and after fine-tuning on our dataset in (a). We present the Ground Truth skeleton and skinning results of the mesh and the results transferred to 3D Gaussian in (b).

can automatically and efficiently produce high-quality rigging results on 3D human characters with rich structures (*e.g., clothing, hair*) originating from multi-modal inputs. In the following, we describe how we preprocess different inputs in Sec. 4.1, which are uniformly converted to a canonical input consisting of 3D Gaussian points and a set of images rendered from 4 views. Then in Sec. 4.2, we sequentially go through our modules for predicting joints, connecting bones, and assigning skinning weights. Finally, tailored for the image inputs, in Sec. 4.3 we propose a simple yet effective module to boost the rendering quality of the rigged characters

4.1. Input Pre-processing

Our framework is flexible in accommodating various input modalities, including text prompts, and anime images. This is in contrast to the prior works [58, 59], which in general are limited to T-pose meshes. In the following, we briefly describe the pre-processing procedure tailored for each modality and refer the readers to Supp. Mat. for more details. As shown in Sec. 3, we have fine-tuned the LGM model on our AnimeRig dataset, which allows for generating faithful 3D Gaussian with a single image of a T-pose character from the frontal view. In the following, we describe how we obtain the desired image input given the following non-3D inputs:

Anime Image: We convert non-T-pose anime images to T-pose using our fine-tuned Animagine-XL [2] model with IP-Adapter [65] and OpenPose [5] ControlNet [68]. This enables standardized T-pose anime image input, allowing the LGM model to generate accurate 3D Gaussian.

Text Prompt: We fine-tune Animagine-XL model with T-pose images obtained from our AnimeRig dataset, which can generate a frontal image of a T-pose character that matches the input text description.

After generating a 3D Gaussian, we extract 3D Gaussian points from the means and render images from four canonical views: front, back, left, and right.

4.2. Rigging for Gaussian Character

Our rigging pipeline is divided into two parts as shown in Fig. 4. In Sec. 4.2.1, we introduce how to predict the positions of joints using a diffusion model based on both geometric and visual cues. We also determine the connections between unordered joints through Minimum Spanning Tre (MST). Finally, we estimate the corresponding skinning with respect to the predicted joint positions in Sec. 4.2.2.

In particular, We train the pipeline with the rigged 3D Gaussian from our proposed AnimeRig dataset. Given a 3D Gaussian, hereafter we denote by μ the means, and by $\{I_i\}_{i=1}^4$ the images rendered from 4 canonical views.

4.2.1. Skeleton Generation

A key challenge in rigging 3D human characters with varying hair and clothing styles arises from the diversity in joint distributions. For instance, a female character in an intricate dress requires more joints to drive than a male character in swimming trunks. The skeletal heterogeneity prevents one from learning a direct mapping from input geometry to joints. Prior works [28, 59] then take an indirect approach, which learns per-point features on input geometry with graph neural networks, and then leverages a non-learnable clustering step to contract surface points towards the labeled joints with regression loss as guidance. On the other hand, since garments are often spatially close to the body, mesh connectivity (*i.e.,* topology) is of critical importance in disentangling their features. The above methods therefore depend on high-quality mesh, which is hard to guarantee in our task of interest.

In response to the lack of topology in 3D Gaussian representation, we propose a novel module, **GSDiff**, a 3D **G**aussian **S**platting conditioned **Diff**usion model for estimating heterogeneous skeletons on our generated 3D Gaussians. To fully exploit 3D Gaussian, **GSDiff** takes the means μ and multi-view rendering $\{I_i\}_{i=1}^4$ of a 3D Gaussian as input, which carry respectively the geometric and visual information. More specifically, we aim to learn to denoise for the joints, $\mathbf{J}$, conditioned on the above input.

We start by introducing a plain diffusion model to learn the joint distribution conditioned on 3D Gaussian input $q(\mathbf{J}|\mu, \{I_i\}_{i=1}^4)$. We can sample from $q(\mathbf{J}|\mu, \{I_i\}_{i=1}^4)$ by starting with noise input and then iteratively sampling from $q(\mathbf{J}_{t-1}|\mathbf{J}_t, \mu, \{I_i\}_{i=1}^4))$. During the reverse process, we use a network s_θ to approximate the conditional distribution:

$$s_\theta(\mathbf{J}_{t-1}|\mathbf{J}_t, \mu, \{I_i\}_{i=1}^4) \approx q(\mathbf{J}_{t-1}|\mathbf{J}_t, \mu, \{I_i\}_{i=1}^4). \quad (1)$$

[2] https://huggingface.co/cagliostrolab/animagine-xl-3.1

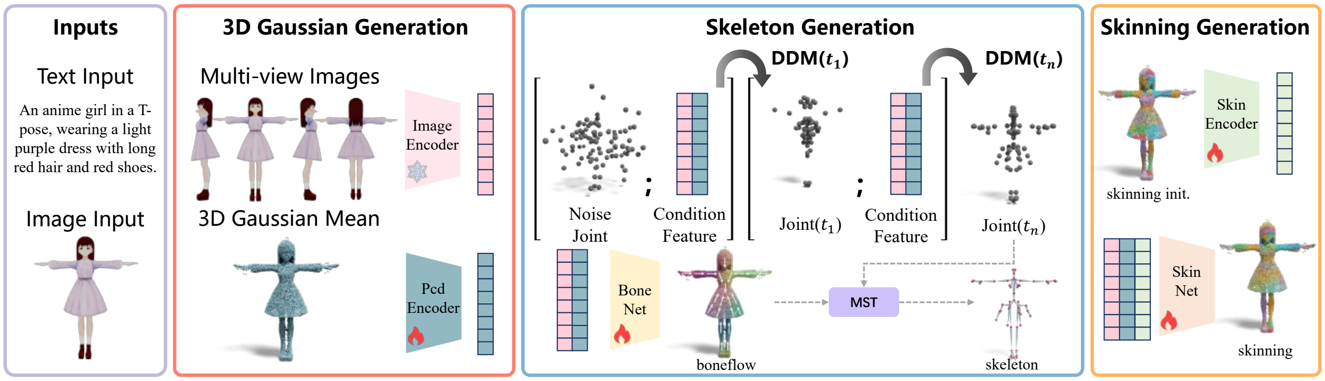

Figure 4. The overall pipeline of our framework. See the main text for more details.

We train a DGCNN [34] as $\mathbf{DG}$ and obtain features for μ. We define the i^{th} joint in the step t as $\mathbf{J}_t^{(i)} \in \mathbb{R}^{1\times 3}$.

Though conditional diffusion has been widely used in various tasks, the denoising object and the conditional input are often comparable (either of the same modality [39] or can be associated by simple operations like rendering [29, 48]). Our task, on the other hand, falls into a new category. Although both μ and $\mathbf{J}$ are point clouds, the latter is effectively an *abstract* of the former, the associating operation is exactly what we seek at the beginning.

We address this problem by proposing a novel conditional diffusion scheme, which enhances the interaction between the conditional inputs and the intermediate denoising outcome. More concretely, considering $\mathbf{J}_t^{(i)}$, the $i-$th joint at the $t-$th step of denoising process. We search for the $k-$nearest neighbors of it among μ, and denote by $\{id_{t,i}^{(1)}, \dots, id_{t,i}^{(k)}\}$ the regarding indices. Then we compute the geometric feature of $\mathbf{J}_t^{(i)}$, $F(\mathbf{J}_t^{(i)})$ as follows:

$$F(\mathbf{J}_t^{(i)}) = \frac{\sum_{j=1}^{k} w_j \mathbf{DG}(\mu_{id_{t,i}}^{(j)})}{\sum_{j=1}^{k} w_j}, \tag{2}$$

where $id_{t,i}^{(l)}$ represents the index of the l^{th} nearest neighbor, $w_j = \frac{1}{\|\mathbf{J}_t^{(i)} - \mu_{id_{t,i}}^{(j)}\|}$ and $\|.\|$ denotes the Euclidean distance. This way we can get $F(\mathbf{J}_t) \in \mathbb{R}^{m\times 128}$ for m joints which integrates information from the surrounding 3D Gaussian, making it easier to determine the position of each joint within the 3D Gaussian accurately.

Moreover, we use CLIP [38] to get the appearance condition feature based on the multi-view images $\{I_i\}_{i=1}^4$. We combine the geometry and appearance feature with the joint feature as the input to each denoising step. We can sample the joint $\mathbf{J}$ by iteratively sample from $s_\theta(\mathbf{J}_{t-1}|F(\mathbf{J}_t), \mathbf{CLIP}(\{I_i\}_{i=1}^4))$. Additionally, inspired by [35], we add a cross-attention layer after each self-attention layer. In particular, we use a modified set transformer as the backbone of the diffusion model [48].

Last but not least, thanks to the rich semantic labels in `AnimeRig`, we separate a set of 25 joints corresponding to body parts from each character. Therefore, $\mathbf{J}$ is further divided into two parts: body joints $\mathbf{J}_b$, and otherwise $\mathbf{J}_o$, which includes hair, clothing, and other un-common parts specific to some character. For ease of learning, we train two separated diffusion models to predict $\mathbf{J}_b$ and $\mathbf{J}_o$ respectively using the same condition.

Bone Connection: Based on our denoised joints, we follow TARig [28] to train a BoneFlow for assisting bone connection. To construct BoneFlow on ground-truth joints, we find for each surface point p its nearest neighbor in joints and the regarding parent. Then BoneFlow at p is defined as the vector pointing from its nearest neighbor to its parent joint. Similarly, we set μ and $\{I_i\}_{i=1}^4$ as input to a network learning to predict BoneFlow. Then we can perform MST algorithm [36] on a cost matrix built on BoneFlow and our estimated joints to obtain bone connection. We refer the reader to the Supp. Mat. for the implementation details.

4.2.2. Skinning Generation

The final step is to estimate the skinning based on the predicted skeleton. Prior arts depend on geometric cues (*e.g.,* geodesic distances [59]) for skinning. Unfortunately, such is unavailable for 3D Gaussian.

However, thanks to our accurate estimation of the skeleton in the last section, we empirically observe that performing k-NN searches between the joints and the 3D Gaussian and constructing $\mathbf{S}_{init} \in \mathbb{R}^{n\times m}$ as the distance-based similarity matrix already yields a decent estimation of skinning. Here n, m are respectively the number of 3D Gaussian points and joints. We defer the details to the Supp. Mat. We then use $\mathbf{S}_{init}$ as the initial estimate for skinning and combine it with 3D Gaussian features to predict the ground truth skinning.

$$\hat{\mathbf{S}} = f_s(\mu, \{I_i\}_{i=1}^4, \mathbf{S}_{init}; \mathbf{W}_s), \tag{3}$$

where $\mathbf{W}_s$ denotes the learned parameters of the skinning estimation.

By treating the per-vertex skinning weights as probability distributions, we use cross-entropy ($\mathcal{L}_{ce}$) and Kullback-Leibler divergence ($\mathcal{L}_{kl}$) as the loss function to measure the disagreement between the ground truth and predicted distributions for each vertex. Since we aim to animate the dense 3D Gaussian, the generated skinning needs to change smoothly to prevent inconsistencies in skinning between adjacent joints, which could cause cracks during the animation. So we introduce a smooth loss as a regularization term.

$$\mathcal{L}_{smooth} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\mathbb{D}_i|}\sum_{j\in\mathbb{D}_i}\left\|\hat{\mathbf{S}}_i - \hat{\mathbf{S}}_j\right\|_2, \tag{4}$$

$$\mathcal{L}_{skinning} = \mathcal{L}_{ce} + \mathcal{L}_{kl} + \lambda_2\mathcal{L}_{smooth}, \tag{5}$$

where $\hat{\mathbf{S}}_i$ denotes the predicted skinning of vertex i, $\mathbb{D}_i$ represents the set of Gaussian points surrounding vertex i and n is the number of 3D Gaussian vertices.

4.3. 3D Gaussian Refinement

In this step, we use a T-pose anime image as a condition to generate 3D Gaussian points. To address the severe artifacts and loss of detail in the head region observed in the original LGM results (Fig. 5(a)) when using a full-body image as input, we fine-tune LGM on our `AnimeRig` dataset and perform separate reconstructions for the head and body. Using SV3D [49] to enhance consistency, we generate new view images at 15° intervals horizontally and select four images at 90° intervals, including the frontal view, as input for LGM. The 3D Gaussian points for the head and body are initially aligned using a fixed cropping frame, refined with the Iterative Closest Point (ICP) algorithm [8], and merged into a single representation. Overlapping points between the head and body Gaussians are filtered out to ensure a smooth transition. Fig. 5(b) shows the results before and after ICP refinement, alongside a comparison with CharacterGen, highlighting the improved alignment and detail preservation achieved by our method. Further details about the refinement stage are provided in the Supp. Mat.

5. Experimental Results

5.1. Metircs and Baselines

To evaluate the predicted skeletons and skinning maps, we adopt the same metrics as RigNet [59]. For skeleton evaluation, we use CD-J2J (The Chamfer distance between joints), CD-J2B (The Chamfer distance between joints and bones), CD-B2B (The Chamfer distance between bones), IoU (Intersection over Union), and Precision & Recall. For skinning evaluation, we leverage Precision & Recall and L1-norm between predicted and reference skinning weights. For more details, please refer to the Supp. Mat.

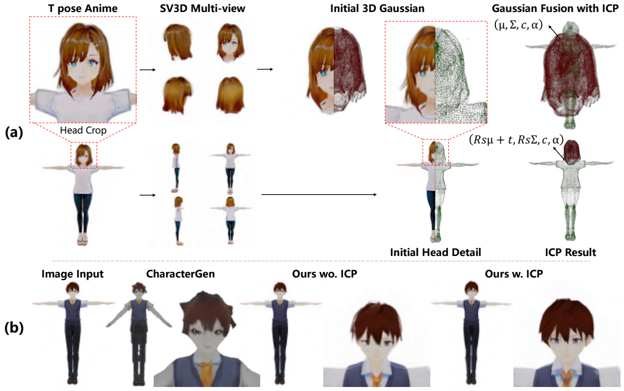

Figure 5. Pipeline for 3D Gaussian refinement and results using a T-pose anime image.

Skeleton prediction: We compare our method with (1) AnimSkelVolNet [58] w. Ground Truth Mesh; (2) RigNet [59] w. Ground Truth Mesh; (3) RigNet w. Mesh from CharacterGen [31]. Specifically, the first two baselines use Ground Truth mesh as input to predict skeleton results with AnimSkelVolNet and RigNet. For a fair comparison, we also apply CharacterGen [31] to produce high-quality 3D mesh predictions from a single character image, which we then use with RigNet for skeleton prediction (the third baseline). Note that AnimSkelVolNet and RigNet are trained on the same dataset as our method.

Skinning prediction: We compare our method with (1) GeoVoxel [9] w. Ground Truth Mesh; (2) RigNet w. Ground Truth Mesh. For GeoVoxel, we utilize the implementation provided by Maya [1]. For all methods, we follow RigNet by using the Ground Truth skeleton as input to predict skinning during training and testing.

5.2. Evaluation for Skeleton and Skinning

Skeleton evaluation: Tab. 1 presents the quantitative comparison in joint estimation. Our results significantly outperform AnimSkelVolNet and RigNet across all metrics. Specifically, the IoU metric, which measures the quality of joint estimation, shows a 59.3% improvement, while the CD-B2B metric, which evaluates the accuracy of bone estimation, shows a 39.9% improvement. For qualitative evaluation, we select examples featuring a variety of genders, hairstyles, and clothing styles, as shown in Fig. 6. RigNet encounters difficulties in accurately estimating skeleton positions, particularly for clothing and hair regions. When using meshes predicted from a single image, challenges such as low mesh quality can even lead to joint estimation failures, such as predicting only one leg. In contrast, our method generates significantly more plausible skeletons based on the predicted 3D Gaussians from a single image. While training RigNet on our dataset, we observe that both the vertex attention network and the final joint prediction network struggle to effectively converge with the more complex skeletal structure. This further highlights the advantages of our approach over regression-based methods.

	IoU ↑	Prec. ↑	Rec. ↑	CD-J2J ↓	CD-J2B ↓	CD-B2B ↓
AnimSkelVolNet	27.74%	29.86%	28.34%	6.45%	4.55%	3.87%
RigNet	28.69%	23.81%	38.13%	5.37%	3.72%	3.26%
Ours	**70.48%**	**70.29%**	**71.91%**	**2.81%**	**2.17%**	**1.96%**

Table 1. Joint prediction results on the test set. AnimSkelVolNet and RigNet are results based on the Ground Truth Mesh.

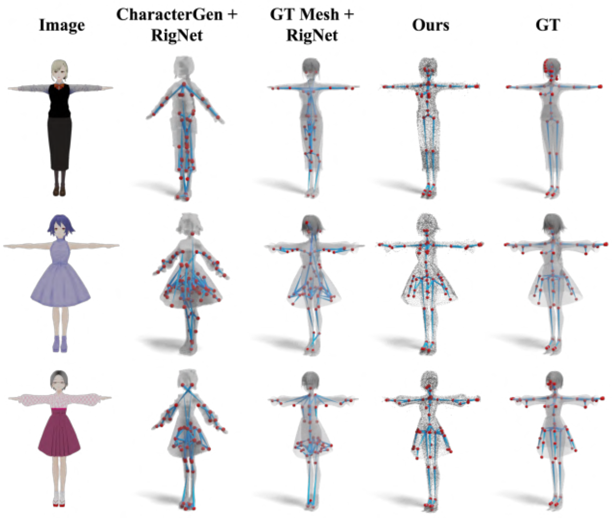

Figure 6. Our method accurately predicts skeletal structures, outperforming RigNet [59] in joint and bone estimation.

Skinning evaluation: Tab. 2 presents the evaluation metrics for skinning. Across all metrics, our results significantly outperform GeoVoxel and RigNet, with a 43.2% improvement in precision and a 45.5% reduction in average L1 error. Note that, due to the non-watertight nature of the mesh, the geodesic distance for RigNet cannot be calculated, so we use Euclidean distance instead. RigNet's low recall in skinning prediction indicates its limitations in accurately predicting control points. Fig. 7 provides a qualitative comparison between our method and the baselines. Our results align more closely with the ground truth overall. While GeoVoxel shows reasonable alignment with the Ground Truth in larger regions, it struggles to capture finer details, and RigNet often fails to produce reasonable results. These qualitative observations are consistent with the quantitative evaluation.

	Prec. ↑	Rec. ↑	avg L1 ↓
GT Mesh w. GeoVoxel	44.48%	69.60%	0.88
GT Mesh w. RigNet	41.94%	35.89%	1.00
Ours	**78.34%**	**71.66%**	**0.48**

Table 2. Skinning prediction results on the test set.

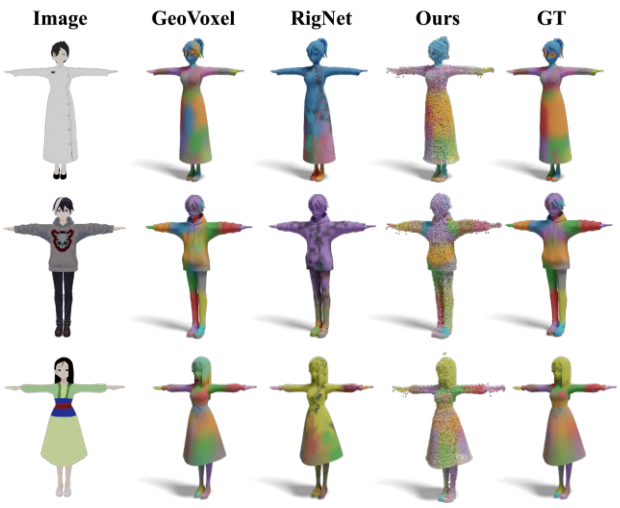

Figure 7. Our skinning predictions closely match the ground truth, surpassing Geovoxel [9] and RigNet [59].

	IoU ↑	Prec. ↑	Rec. ↑	CD-J2J ↓	CD-J2B ↓	CD-B2B ↓
Regression	46.72%	47.14%	47.05%	4.02%	3.07%	2.59%
Ours w/o C_{3d}	56.56%	56.18%	58.12%	3.36%	2.29%	2.39%
Ours w/o C_{3dl}	58.71%	59.18%	59.49%	4.14%	3.32%	2.82%
Ours w/o C_i	67.14%	68.02%	67.53%	2.96%	2.55%	2.24%
Ours	**70.48%**	**70.29%**	**71.91%**	**2.81%**	**2.17%**	**1.96%**

Table 3. Ablation study on joint estimation. C_{3d} denotes 3D Gaussian condition, C_{3dl} denotes 3D Gaussian local condition, and C_i denotes image condition.

5.3. Ablation study

Generation beats Regression: Here, we discuss the performance of joint position estimation using regression-based and generation-based approaches. As a baseline, we implement a regression-based method for estimating joint positions by replacing RigNet's mesh encoder with the current state-of-the-art point cloud encoder [51] and creating a pipeline similar to RigNet for regressing joint positions from 3D Gaussians. We train and test this model on the same dataset, with the quantitative results for joint prediction presented in the first row of Tab. 3. Our observations show that the diffusion-based method significantly outperforms the regression-based approach in joint estimation. Regression methods face challenges in learning the positional information of skeletons with varying topologies, whereas diffusion-based methods, by learning positional distributions, exhibit a clear advantage.

Conditional generation: We discuss the impact of different conditioning methods on the generated results during the diffusion-based joint generation process, with outcomes presented in Tab. 3. First, we conduct an ablation study without any 3D information (only image inputs), resulting in a 19.8% drop in the IoU metric. This empha-

	IoU ↑	Prec. ↑	Rec. ↑	CD-J2J ↓	CD-J2B ↓	CD-B2B ↓
Plain Mesh	67.90%	69.40%	67.73%	3.02%	2.31%	2.04%
Textured Mesh	71.05%	70.78%	72.32%	2.38%	1.87%	1.71%
3D Gaussian	70.48%	70.29%	71.91%	2.81%	2.17%	1.96%

Table 4. Quantitative results of joints prediction using mesh and 3D Gaussian as input.

sizes the importance of incorporating geometric information from the 3D Gaussian in joint position estimation. We then test without the 3D Gaussian local condition that directly uses the 3D Gaussian point global feature as input, leading to a 16.7% decrease in the IoU metric. This result suggests that local information from the 3D Gaussian is essential for accurately associating with joint data. Finally, when we remove the image input, we observe a slight decrease in performance, indicating that appearance information also contributes to joint estimation. Overall, these ablation experiments validate the effectiveness of each component in our conditioning approach.

We experiment with replacing the input of **GSDiff** with mesh, using mesh as input for both training and testing. As shown in Tab. 4, both inputs yield similar results, demonstrating the flexibility of our method.

Skinning generation: We also conduct experiments on the impact of different inputs for skinning estimation as shown in Tab. 5. First, we remove the rendered image input from 3D Gaussian, and the final average L1 error increased by 30.4%, indicating that the appearance information from 3D Gaussian is crucial for skinning learning. We then experiment by removing the initial skinning values, which also led to an increase in average L1 error of 26.2%, demonstrating that S_{init} is essential for the skinning learning process. Finally, we show that the smoothness of 3D Gaussian skinning learning can also help the network converge.

	Prec. ↑	Rec. ↑	avg L1 ↓
Regression	39.46%	33.93%	1.32
Ours w/o image input	62.93%	71.01%	0.69
Ours w/o S_{init} input	57.46%	70.48%	0.65
Ours w/o smooth loss	77.36%	71.31%	0.49
Ours	**78.34%**	**71.66%**	**0.48**

Table 5. Ablation study on skinning prediction. S_{init} denotes initial skeleton input.

5.4. Applications

Rigging with Any Pose Anime Image Inputs: In this section, we show an application of rigging from an arbitrary pose anime image to the final rigging results. We select image data from [31], which has a certain domain gap compared to our dataset. We first consistently convert the input image to a T-pose, then reconstruct the 3D Gaussian, and finally generate the corresponding rigging results, as shown in Fig. 8. Our method successfully produces reasonable skeletons and skinning for various clothing and hairstyles, demonstrating the generalization capability and applicability of our approach.

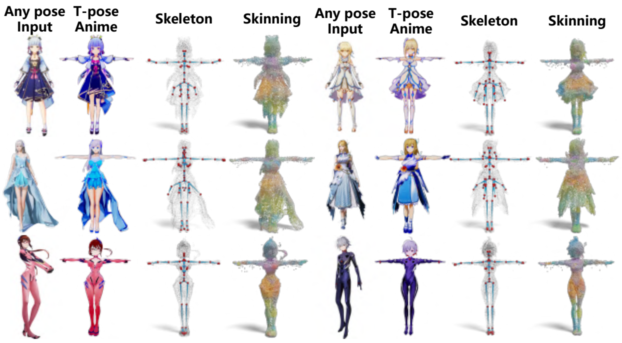

Figure 8. Our pipeline can consistently convert anime characters in any pose to T-pose and provide a reliable rigging result.

Animation for rigged characters: Here, we present the results of animating the rigged 3D Gaussian model as shown in Fig. 2. Compared with CharacterGen, our 3D Gaussian-based method produces higher-resolution rendered results. Additionally, the generated skeletons for the skirt and hair display more natural movement, effectively preventing issues like the skirt and legs moving in unison.

6. Conclusion, Limitation, and Future Work

In this work, we introduce **DRiVE**, the first framework for generating and rigging 3D characters with rich structures using 3D Gaussian representations. Our approach achieves state-of-the-art accuracy in skeleton and skinning predictions, enabling detailed modeling of hair and garments for realistic animations. By curating a large-scale rigged character dataset, `AnimeRig`, and proposing the novel diffusion module **GSDiff**, we effectively address the challenges of data insufficiency and skeleton prediction on 3D Gaussian. The flexibility of our framework allows it to be extended beyond 3D Gaussian to mesh, further demonstrating its versatility. We believe DRiVE offers valuable insights for rigging and animation, with promising potential for practical applications.

We also identify the following limitations, which lead to future work directions: 1) The node density of our generated skeleton is deterministic, and cannot be tuned according to user preferences. It would be interesting to learn from more versatile data; 2) Unlike meshes, there is no mature solution for 3D Gaussian collision detection. Thus our rigged 3D Gaussian can suffer from self-crossing among parts (*e.g.,* leg and dress). While the high rendering frame rate allows for efficient human inspection, it would be interesting to explore more automatic solutions.

References

[1] Autodesk 2019 maya version. www.autodesk.com/products/autodesk-maya/. 2019. 6

[2] Ilya Baran and Jovan Popović. Automatic rigging and animation of 3d characters. *ACM Transactions on graphics (TOG)*, 26(3):72–es, 2007. 2, 3

[3] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint arXiv:2408.00653*, 2024. 3

[4] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 958–968, 2024. 3

[5] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 4

[6] Junhao Chen, Xiang Li, Xiaojun Ye, Chao Li, Zhaoxin Fan, and Hao Zhao. Idea-2-3d: Collaborative lmm agents enable 3d model generation from interleaved multimodal inputs. *arXiv preprint arXiv:2404.04363*, 2024. 3

[7] Mingjin Chen, Junhao Chen, Xiaojun Ye, Huan-ang Gao, Xiaoxue Chen, Zhaoxin Fan, and Hao Zhao. Ultraman: Single image 3d human reconstruction with ultra speed and detail. *arXiv preprint arXiv:2403.12028*, 2024. 3

[8] Dmitry Chetverikov, Dmitry Svirko, Dmitry Stepanov, and Pavel Krsek. The trimmed iterative closest point algorithm. In *2002 International Conference on Pattern Recognition*, pages 545–548. IEEE, 2002. 3, 6

[9] Olivier Dionne and Martin de Lasa. Geodesic voxel binding for production character meshes. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 173–180, 2013. 3, 6, 7

[10] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7792–7801, 2019. 3

[11] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–549, 2024. 3

[12] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[13] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 634–644, 2024. 1, 3

[14] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9352–9364, 2023. 3

[15] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2024. 1

[16] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*, 2024. 3

[17] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 3

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2

[19] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024. 3

[20] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 3

[21] Lijuan Liu, Youyi Zheng, Di Tang, Yi Yuan, Changjie Fan, and Kun Zhou. Neuroskinning: Automatic skin binding for production characters with deep graph networks. *ACM Transactions on Graphics (ToG)*, 38(4):1–12, 2019. 3

[22] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[23] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 3

[24] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. *arXiv preprint arXiv:2311.17061*, 2023. 3

[25] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[26] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang,

Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 3

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3

[28] Jing Ma and Dongliang Zhang. Tarig: Adaptive template-aware neural rigging for humanoid characters. *Computers & Graphics*, 114:158–167, 2023. 3, 4, 5

[29] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12923–12932, 2023. 5

[30] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3

[31] Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. *ACM Transactions on Graphics (TOG)*, 43(4): 1–13, 2024. 1, 2, 3, 6, 8

[32] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 3

[33] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 3

[34] Anh Viet Phan, Minh Le Nguyen, Yen Lam Hoang Nguyen, and Lam Thu Bui. Dgcnn: A convolutional neural network over large-scale labeled graphs. *Neural Networks*, 108:533–543, 2018. 5

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[36] Robert Clay Prim. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6): 1389–1401, 1957. 5

[37] Hongxing Qin, Songshan Zhang, Qihuang Liu, Li Chen, and Baoquan Chen. Pointskelcnn: Deep learning-based 3d human skeleton extraction from point clouds. In *Computer Graphics Forum*, pages 363–374. Wiley Online Library, 2020. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 5

[40] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 1, 3

[41] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 84–93, 2020. 1, 3

[42] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. *arXiv preprint arXiv:2403.05087*, 2024. 3

[43] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[44] Zhaoqi Su, Liangxiao Hu, Siyou Lin, Hongwen Zhang, Shengping Zhang, Justus Thies, and Yebin Liu. Caphy: Capturing physical properties for animatable human avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14150–14160, 2023. 3

[45] Jeff Tan, Gengshan Yang, and Deva Ramanan. Distilling neural fields for real-time articulated shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4692–4701, 2023. 3

[46] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 2, 3

[47] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 3

[48] Michał J Tyszkiewicz, Pascal Fua, and Eduard Trulls. Gecco: Geometrically-conditioned point diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2128–2138, 2023. 5

[49] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. 6

[50] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 3

[51] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 7

[52] Yuefan Wu, Zeyuan Chen, Shaowei Liu, Zhongzheng Ren, and Shenlong Wang. Casa: Category-agnostic skeletal animal reconstruction. *Advances in Neural Information Processing Systems*, 35:28559–28574, 2022. 3

[53] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 1, 3

[54] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 512–523, 2023. 1

[55] Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitrios Tzionas, and Michael J Black. Puzzleavatar: Assembling 3d avatars from personal albums. *arXiv preprint arXiv:2405.14869*, 2024. 1, 3

[56] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3

[57] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2024. 3

[58] Zhan Xu, Yang Zhou, Evangelos Kalogerakis, and Karan Singh. Predicting animation skeletons for 3d articulated models via volumetric nets. In *2019 international conference on 3D vision (3DV)*, pages 298–307. IEEE, 2019. 2, 3, 4, 6

[59] Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. Rignet: Neural rigging for articulated characters. *arXiv preprint arXiv:2005.00559*, 2020. 2, 3, 4, 5, 6, 7

[60] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Jiashi Feng, and Mike Zheng Shou. Xagen: 3d expressive human avatars generation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[61] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022. 3

[62] Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16995–17005, 2023. 3

[63] Ji Yang, Xinxin Zuo, Sen Wang, Zhenbo Yu, Xingyu Li, Bingbing Ni, Minglun Gong, and Li Cheng. Object wakeup: 3d object rigging from a single image. In *European Conference on Computer Vision*, pages 311–327. Springer, 2022. 3

[64] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13284–13293, 2021. 3

[65] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *preprint arXiv:2308.06721*, 2023. 4

[66] Xianfang Zeng, Xin Chen, Zhongqi Qi, Wen Liu, Zibo Zhao, Zhibin Wang, Bin Fu, Yong Liu, and Gang Yu. Paint3d: Paint anything 3d with lighting-less texture diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4252–4262, 2024. 2

[67] Hao Zhang, Fang Li, Samyak Rawlekar, and Narendra Ahuja. S3o: A dual-phase approach for reconstructing dynamic shape and skeleton of articulated objects from single monocular video. *arXiv preprint arXiv:2405.12607*, 2024. 3

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4

[69] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 3

[70] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9947, 2024. 3