Joint2Human: High-quality 3D Human Generation via Compact Spherical Embedding of 3D Joints

Muxin Zhang[1,†], Qiao Feng[1,†], Zhuo Su[2], Chao Wen [2], Zhou Xue[3], Kun Li[1,*]
[1]College of Intelligence and Computing, Tianjin University
[2]PICO IDL, ByteDance [3]Li Auto

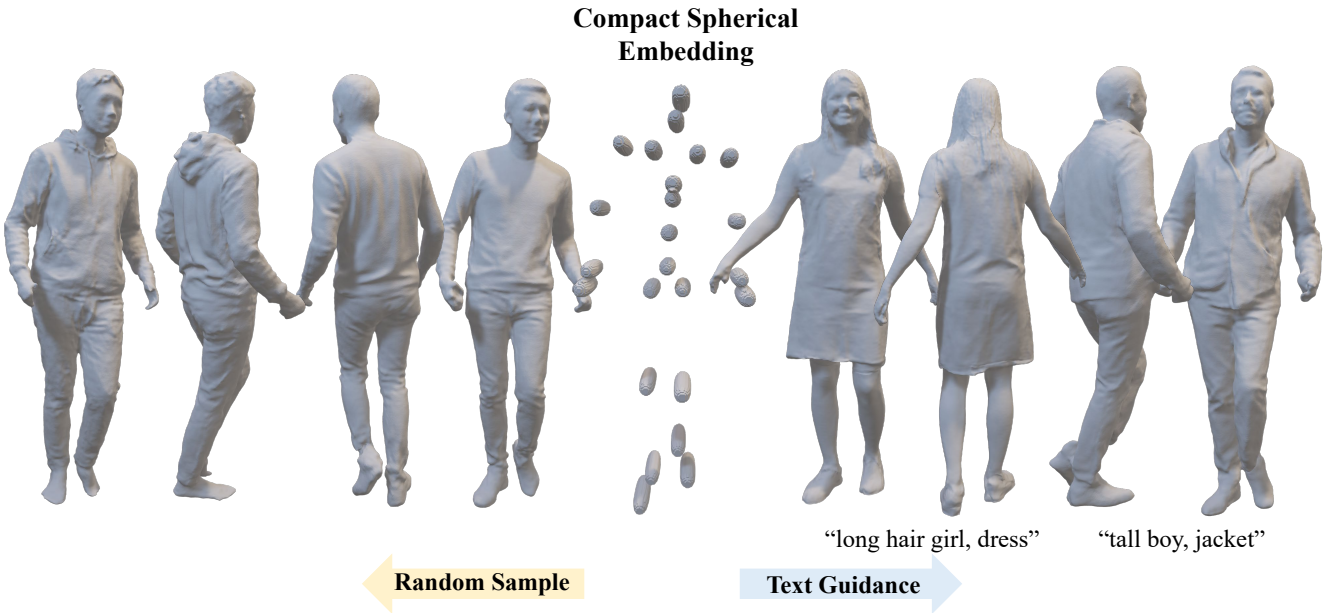

Figure 1. With the proposed compact spherical embedding of 3D joints, our method can generate high-quality and high-resolution 3D humans with reasonable global structure and fine-grained geometry details in an efficient way, based on a native conditional 3D generative network with 2D diffusion model.

Abstract

3D human generation is increasingly significant in various applications. However, the direct use of 2D generative methods in 3D generation often results in losing local details, while methods that reconstruct geometry from generated images struggle with global view consistency. In this work, we introduce Joint2Human, a novel method that leverages 2D diffusion models to generate detailed 3D human geometry directly, ensuring both global structure and local details. To achieve this, we employ the Fourier occupancy field (FOF) representation, enabling the direct generation of 3D shapes as preliminary results with 2D generative models. With the proposed high-frequency enhancer and the multi-view recarving strategy, our method can seamlessly integrate the details from different views into a uniform global shape. To better utilize the 3D human prior and enhance control over the generated geometry, we introduce a compact spherical embedding of 3D joints. This allows for an effective guidance of pose during the generation process. Additionally, our method can generate 3D humans guided by textual inputs. Our experimental results demonstrate the capability of our method to ensure global structure, local details, high resolution, and low computational cost simultaneously. More results and the code can be found on our project page at http://cic.tju.edu.cn/faculty/likun/projects/Joint2Human.

† Equal contribution.
* Corresponding author.

1. Introduction

The field of 3D human generation holds considerable significance across diverse applications, including virtual/augmented reality, gaming, and the film industry. Despite advancements in existing methods [9, 22, 46], it remains a challenge to achieve a simultaneous guarantee of both global structural accuracy and local geometry details, which is often compounded by high computational costs. In this paper, we aim to address these limitations and propose a method for generating high-quality 3D human models that exhibit fidelity to both global structures and local details while ensuring computational efficiency, as illustrated in Figure 1.

Existing human generation methods can be divided into two categories: two-stage methods and native 3D generation methods. Two-stage generation methods [19, 20, 22, 35, 46] fit 3D humans from 2D images via NeRF [31] or differentiable rendering [32]. These methods are trained only on 2D datasets and hence lack a 3D backbone for perceiving 3D structures and ensuring view consistency. Chupa [22] aims to produce high-quality results through dual normal map-based optimization but suffers from depth ambiguity issues that harm view consistency. Most native 3D generation methods [1, 9, 34, 44] generate 3D humans directly with Tri-planes [4] or Signed Distance Fields [35]. However, these approaches encounter challenges in achieving high-fidelity global structures with fine-grained local geometry details. Additionally, their generalization capabilities are limited, and computational efficiency is compromised. In summary, the existing methodologies collectively fail to ensure global structural fidelity, local detail preservation, high resolution, and computational efficiency simultaneously.

To efficiently generate high-quality 3D humans with reasonable global structure and fine-grained geometry details, in this paper, we propose *Joint2Human*, a conditional generative network with 2D diffusion models derived from 3D datasets. To achieve high-fidelity 3D human generation with reasonable global structure, we employ the 2D generative models to produce 3D shapes as preliminary results directly. Subsequently, we carefully design the compact spherical embedding of 3D joints based on an image-aligned 3D representation FOF [10, 11]. With it, we implement effective pose guidance and diverse generation. We also design a high-frequency enhancer and a multi-view recarving strategy for fine-grained local detail generation. Experimental results demonstrate that our method outperforms the state-of-the-art methods regarding global structure, local detail, and computational efficiency. Furthermore, our method also exhibits versatility by enabling the generation of 3D human representations guided by text.

To summarize, our main contributions include:

- We propose *Joint2Human*, a native conditional 3D generative method with a 2D diffusion model for high-quality, high-resolution 3D human generation. To our knowledge, it is the first work based on the FOF that can simultaneously ensure global structure, local details, high resolution, and low computational cost.
- We propose a new pose guidance embedding, a compact spherical embedding of 3D human joints, for efficient perception of global structure. This mechanism also facilitates a more straightforward and effective implementation of pose-guided generation in 2D generation framework.
- We design a high-frequency enhancer by integrating a subsidiary decoder into the pre-trained VAE and a multi-view recarving strategy for fine-grained local detail generation. Both of them improve the geometry quality of the final results.

2. Related Work

2.1. 3D Human Generation with 2D Generators

Many approaches [12, 19, 20, 22, 35, 46] try to learn the 3D shape from 2D images via various NeRF representations [25, 28, 31, 33, 38, 39] and differentiable volume rendering [32, 45, 51]. However, it is always computationally expensive and limited in resolution. EVA3d [19] designs complex training strategies to achieve high-resolution generation, but its geometric quality is mediocre. What's worse, the 2D dataset is always imbalanced in viewing angles and human poses. Hence, these methods are relatively unstable, and it is hard to generate a realistic full-body human geometry. Other methods [20, 22, 46] also leverage the priors from 2D human generation and 3D human reconstruction models, which is cumbersome. In Chupa [22], many optimizations are needed to maintain view consistency. It first generates two normal maps for a clothed human's front and back sides, then optimizes the 3D generated mesh by the dual normal maps. There are also some human template-based works [3, 18, 22, 27, 48, 50] that are highly dependent on human template mesh, such as SMPL [29] and SMPL-X [37], which adversely affect the diversity of generated models especially in local details. For example, it is flawed in modeling loose clothes like dresses.

2.2. 3D-aware Generation

Based on the development of 3D human representation, numerous methods have emerged for native 3D generation. Some of these approaches [1, 9, 34, 44] utilize tri-planes [4] which is a NeRF-based representation. It is hard to generate precise geometry, which is limited by computational complexity. Apart from that, the implicit functions [35, 36] are highly favored in several methods [3, 18, 19, 24, 46]. Alternatively, some methods use 3D data for training, gDNA [6] use implicit multi-subject forward skinning which enables learning from 3D scans of

Table 1. Comparison with existing methods.

Method	Global Structure (View Consistency)	Local Detail	High Resolution	Loose Clothes	Computational Efficiency
GNARF[1]	✓	✗	✗	✗	✗
AG3D [9]	✓	✗	✗	✗	✗
Get3Dhuman [46]	✗	✓	✓	✓	✗
Chupa [22]	✗	✓	✓	✗	✗
Ours	✓	✓	✓	✓	✓

human bodies. Among all the works mentioned above, the GAN [1, 4, 9, 13, 19, 20, 34, 35, 46, 50] and Diffusion models [3, 17, 22, 24, 44] are the most popular.

Using the diffusion model for small object generation[5, 7, 8, 26] is more straightforward to combine with text control. However, the text-guided 3D human generation is more difficult due to the high-dimensional cube and lack of text-3D data pairs. Some attempts were made to tackle these issues. AvatarCLIP [18] initialize a bare mesh shape with SMPL. It continuously optimizes the mesh to match the input text by calculating the CLIP score between the text token and the mesh's rendering image. CLIP-actor [48] does the same thing, except it designs a recommendation module to get the initial shape/motion. In this way, they reduced the need for text-3D data pairs with the help of the large text-image pre-trained model [40]. Other methods [3, 22, 44] impose textual controls on the diffusion model for generated shapes that match the input description. In previous approaches [7, 8, 26] to generating small objects, the resolution of the generated geometry is limited for human generation, such as 64^3 or 128^3. In addition, some methods, such as differentiable rendering-based methods, are time-consuming to infer.

Different from the two types of methods described above, as shown in Tab. 1. Our approach can accommodate both global structures and local details. To fully perceive the global structures, we adopt the image-aligned 3D representation FOF and introduce a compact spherical embedding of 3D joints for pose guidance. To generate details, we design a high-frequency enhancer and a multi-view recarving strategy in 3D space. Apart from this, we also simultaneously achieve high-resolution generation and low computational cost.

3. Method

Our method aims to generate diverse 3D-clothed humans with global structure and local details. First, we propose a direct conditional 3D generative method with 2D diffusion models as illustrated in Fig. 2. Specifically, we utilize an image-aligned 3D representation FOF to produce 3D shapes directly. We operate the condition and generation process in the latent space (Sec. 3.1). Then, we propose the compact spherical embedding of 3D joints (Sec. 3.2) for precise pose control and diverse generation. Furthermore, a high-frequency enhancer (Sec. 3.3) and a multi-view recarving strategy (Sec. 3.4) are proposed to improve the geometry details further. We will detail all the stages of our method in the following subsections.

3.1. Latent Diffusion for Fourier Occupancy Fields

To model the distribution of Fourier Occupancy Fields, following the latent diffusion model [41], we utilize an auto-encoder to compress the high-dimensional raw data space into a lower-dimensional latent space, which encodes human shapes into a normal distribution. In detail, we adopted a VAE [23], which contains the encoder $\mathcal{E}$ and decoder $\mathcal{D}$. Given a FOF feature $x \in \mathbb{R}^{512\times512\times32}$, the encoder $\mathcal{E}$ encodes FOF into latent vectors $z = \mathcal{E}(x)$, where $z \in \mathbb{R}^{128\times128\times8}$. The decoder $\mathcal{D}$ decodes the latent vectors back to the original FOF space $\tilde{x} = \mathcal{D}(\mathcal{E}(x))$, where $\tilde{x} \in \mathbb{R}^{512\times512\times32}$. We pretrained the auto-encoder with the reconstruction loss and KL-regularization loss like [41]. During training, the training data is passed through the encoder $\mathcal{E}$ in the vae module to obtain its feature in the latent space. We adopt a U-Net-like structure as our denoising model ϵ_θ. We follow the classical and efficient loss function proposed by Ho [17] and train the diffusion model based on the T steps of noise-adding and denoising processes in the latent space. Thus, we can obtain a vector $\tilde{z}_0$ in the latent space, which will be sent to the decoder $\mathcal{D}$ to get FOF. Then, we can convert the FOF to an occupancy field and get an initial 3D human mesh from it with the Marching Cubes [30] algorithm.

3.2. Condition-guided Generation Mechanisms

Compact Spherical Embedding of 3D Joints. Different from previous works [3, 9, 18, 22, 48, 50] adopt the human parametric model [29, 37] for pose guidance, we find that such a strategy with FOF can lead to overfitting. Conditioning on the SMPL makes it harder for networks to learn the true data distribution, which is also mentioned

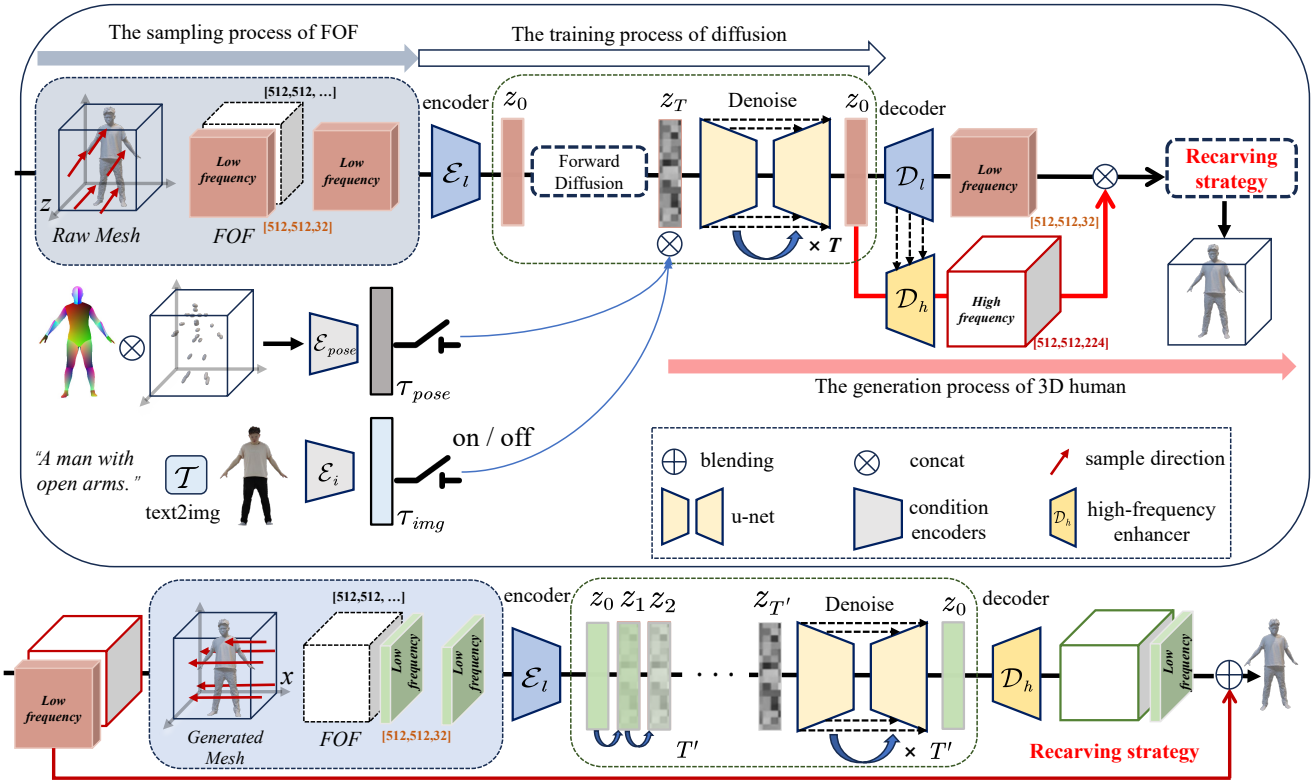

Figure 2. **Overview of our framework.** *Joint2Human* takes Gaussian noise and some conditional information as input. We first encode FOF into a latent space, where a diffusion model is trained to enable high-resolution human generation. Furthermore, to enable flexible conditional generation, we adopt condition encoders along with classifier-free guidance to enable conditional generation. Our conditional control generation strategy can support switching between different modalities. Then, we pass the results generated by the first denoising process through the high-frequency enhancer $\mathcal{D}_h$ and multi-view recarving strategy for fine-grained local detail generation. Additionally, the recaiving strategy is shown in the bottom subfigure.

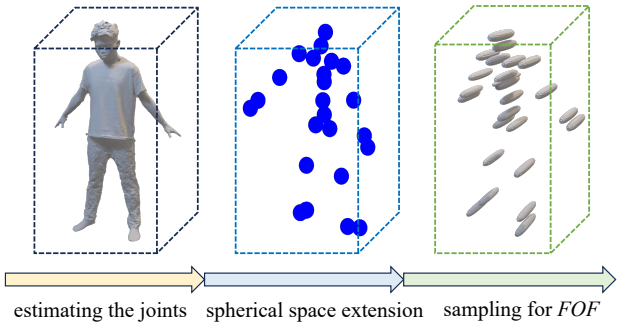

Figure 3. The data processing flow for compact spherical embedding of 3D joints.

in ECON [47]. To avoid the redundant information and the misleading geometric prior from SMPL, we use the human body joints as pose guidance and design a compact spherical embedding of 3D joints. As shown in Fig. 3, we first estimate the K 3D joints location $J = \{p_i \in \mathbb{R}^3\}_{i=1}^{K}$ of the given human. For each joint position p_i, we extend a sphere with p_i as the center and r as the radius in 3D space. After that, according to the Sec. 3.1, we compute the FOF $c_i \in \mathbb{R}^{512\times512\times8}$ for each sphere. Therefore, the depth information is stored in the phase of trigonometric functions. We concatenate $\{c_i\}_{i=1}^{K}$ channel-wise with a fixed order to form the compact spherical embedding of 3D joints $con_{joint} \in \mathbb{R}^{512\times512\times8K}$. In this way, our approach processes joints in different channels and concatenates them to form the embedding of 3D joints, integrating full semantic and depth-wise information for precise pose control. In our setting, we use $r = 10cm$ and follow the joints $K = 24$ in SMPL.

Pose-guided Geometry Generation. We aim to generate a human shape that fits the given pose. With the compact spherical embedding of the 3D joints, we can efficiently perform pose control in human generation. To further improve the stability of the generative model, we incorporate the 2D prior by utilizing the IUV map con_{iuv} defined in

DensePose [14]. During training, we first get the con_{pose} by concatenating the con_{joint} and con_{iuv} in the channel dimension. Then, we learn a pose encoder ε_{pose} to get the human pose embedding τ_{pose} and apply the pose condition con_{pose} on the latent code z by concatenation after passing through the conditional encoder :

$$\tau_{pose} = \varepsilon_{pose}\left(con_{pose}\right), \tag{1}$$

where the ε_{pose} consists of two convolutional layers for the alignment of the feature. In the inference stage, we first get the human pose embedding τ_{pose}; then we perform the denoising process of the ϵ_θ to get the initial human shape.

Text-guided Geometry Generation. To generate the 3D human with the guidance of text description, we proposed the image prompt strategy for our model based on a text-guided 2D human image generation model Text2Human [21]. In the training stage, we learn an additional image encoder ε_i to get the human image embedding τ_{img}, then apply it to the training of the diffusion model by concatenation. In the inference stage, with the input text, we first infer the text-to-image model $\mathcal{T}$ and image encoder ε_i to get the embedding τ_{img}, then we sample the 3D human shape from ϵ_θ with all these conditions. We provide additional text-guided generation results in the supplementary materials.

3.3. High-frequency Enhancer

For efficient human representation and modeling of data distribution, we only save low-frequency terms while taking the FOF as a 3D representation in Sec. 3.1. However, some abandoned high-frequency information is essential for human shape details. Therefore, based on the generation of the low-frequency FOF feature, we propose the High-Frequency enhancer to recover these missing details C_h, as shown in Fig. 2. In detail, we learn a reference-based decoding network $\mathcal{D}_h$ based on the latent space and decoder from the auto-decoder claimed in Sec. 3.1 to predict high-frequency terms $\hat{C}_h$:

$$\hat{C}_h = \mathcal{D}_h\left(\mathcal{D}_l\left(z_0\right), z_0\right), \tag{2}$$

To enhance the perception of latent-space contextual information, we introduce additional skip connections for each layer in $\mathcal{D}_h$, which is used to fuse the feature map $\mathcal{D}_l\left(z_0\right)$ from the relevant layers in the decoder $\mathcal{D}_l$. When training, we calculate the MSE loss between the predicted terms C_h and high-frequency ground truth C_h as the supervision to optimize $\mathcal{D}_h$,the loss function $\mathcal{L}_h$ is formulated as:

$$\mathcal{L}_h = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{C}_h\left(x_i, y_i\right) - C_h\left(x_i, y_i\right)\right)^2, \tag{3}$$

where n is the number of all pixels in C_h. So, in the inference stage, we can estimate the high-frequency feature based on the known low-frequency feature to enhance the geometric details.

3.4. Multi-view Recarving Strategy

Since our training data is sampled along a fixed direction and the directionality of FOF, there are a few artifacts along the direction orthogonal to the normal direction of the generated shape. We propose the multi-view recarving strategy to improve the geometry quality while maintaining muti-view consistency.

Different from the resampling in Chupa [22], we perform this process on 3D space and fuse them by blending the occupancy fields. In detail, we first perform the inference process of diffusion and high-frequency enhancer to obtain the FOF C_{init} and the occupancy field F_{init}. Based on the F_{init}, we can get the initial human mesh $\mathcal{M}_{init}$, which may have some artifacts in other views. To tackle these artifacts, we rotate the $\mathcal{M}_{init}$ θ degrees along the yaw axis to get $\mathcal{M}_\theta$. After that, we convert the $\mathcal{M}_\theta$ into FOF C_θ along the current orientation. With the C_θ, we do the same thing as described in Sec. 3.1: leveraging the encoder $\mathcal{E}$ to get the latent code $\hat{z}$. Then, we re-execute the forward process of diffusion, adding T' steps of noise to $\hat{z}$ and denoising the noise $\hat{z}$ with ϵ_θ. We pass the denoised latent code through the decoder to get the FOF C_θ. Meanwhile, we can reconstruct the occupancy field F_θ from FOF C_θ by efficient Fourier inversion. At last, we blend the occupancy fields $\{F_{init}, F_{\theta_1}, F_{\theta_2}...\}$ of different views and extract the 3D human mesh from the occupancy fields. In our setting, we perform this procedure only once with the $\theta = \frac{\pi}{2}$. In this way, two orthogonal views are blended in a weighted-average approach. We settle on this setup under the trade-off between computational efficiency and geometric quality.

4. Experiments

4.1. Experimental Setup

- **Datasets.** We train our model with THuman 2.0 [49], THuman 3.0 [43], 2k2k [15] and about 1500 high-quality meshes from commercial datasets. To ensure a fair comparison, we deviated from Chupa [22]'s settings and used a third-party dataset CustomHumans [16] as our test set. This dataset wasn't used to train any model. We do this because the Sota methods' training code is not publicly available, and some of the datasets they use are not publicly available. For the training stage, we first sample 32 successive different angles of FOF along the yaw axis of the rotating human body to generate FOF feature maps for the same mesh. After that, we obtain the joint points for each mesh and calculate the compact spherical embedding of the 3D joint using the same FOF sampling process.
- **Baselines.** We compare our method with Chupa [22] and

Table 2. Quantitative evaluation. We report two types of FID scores on the test dataset.

Method	$\text{FID}_{normal} \downarrow$	$\text{FID}_{shade} \downarrow$
Chupa_{coarse}	51.60	73.03
Chupa_{fine}	29.90	45.49
Ours_{coarse}	37.67	59.45
Ours_{fine}	**23.89**	**41.20**

AG3D [9] as baselines. Chupa is the current state-of-the-art method for generating 3D human geometry. We don't make a quantitative comparison with AG3D because it uses 2D datasets, and the training code of AG3D is unavailable.

- **Metrics.** We measure the quality of the generated human mesh by using the Fréchet Inception Distance (FID) [2] between the rendering normal maps [6, 22] and shading images [22, 42] of the generated meshes. We follow the settings of Chupa [22] and render the meshes into 18 views with 20° yaw interval for calculating FID.

4.2. Implementation Details

To represent and generate a high-quality 3D human geometry, the channels of the FOF feature maps need to be set to at least 32. For the training stage of the FOF autoencoder, we train it for three days on 8 NVIDIA A100 GPUs, with a batch size of 32. For the diffusion model training stage, we train it for eight days on 8 NVIDIA A100 GPUs, with a batch size of 64. The total number of time steps is set as $T = 1000, T' = 200$ for the diffusion model in our pipeline. We present more details in the supplementary material.

4.3. Generated Results

Fig. 4 shows various generated results by our method. Our method can generate high-quality 3D humans with global structure, local details, and high resolution. Benefiting from our proposed compact spherical embedding of 3D joints, we can flexibly generate high-quality 3D humans guided by poses as shown in Fig. 5. Our model performs well in the diversity of generations with the fixed pose as guidance. It can generate humans with different identities and costumes.

4.4. Comparison

Qualitative and Quantitative Results. We compare our method with the latest human generative models Chupa [22] and AG3D [9]. In detail, We conduct a quantitative comparison with Chupa to assess the quality of the generated meshes. We also conduct a qualitative comparison with AG3D and Chupa. To ensure a fair comparison, we try to make the input poses of Chupa and our method the same as much as possible. However, AG3D is not friendly to fixed poses of the results, so we used random poses in the test stage. The quantitative results for human generation are shown in Tab. 2. Our method outperforms the current SOTA methods and achieves better FID on both images. Fig. 6 shows qualitative comparison results. Our results gain the natural and detailed visual effects. Besides, we have better view consistency when compared to the generated results of other methods. In addition, our model can complete the generation process in less than one minute. It is advanced in loose clothing modeling and diversity generation. The discussion on running time and the visualized generation results are shown in the supplementary material.

Figure 4. Our method can generate various high-quality 3D humans.

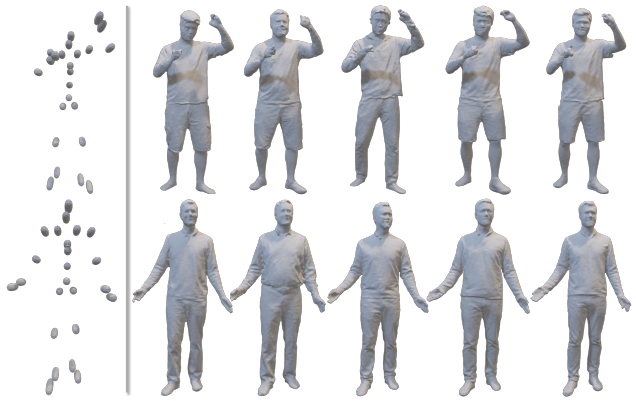

Figure 5. Our method can generate diverse results given a specific pose.

Figure 6. The generated 3D humans compared with Chupa [22] and AG3D [9].

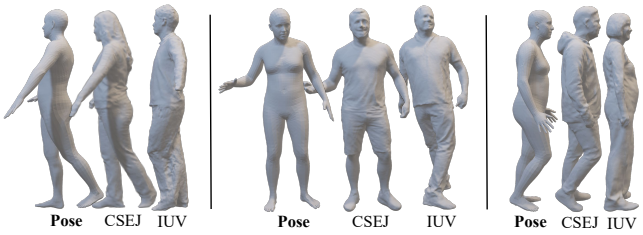

Figure 7. The generated results with different pose-guidance strategies.

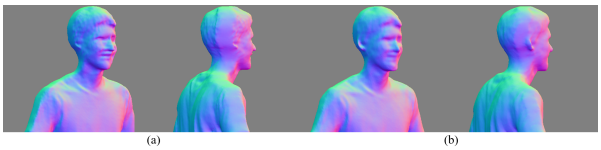

Figure 8. We present the side-view normal maps of the generated human. (a) illustrates the results without the High-Frequency Enhancer, while (b) illustrates the results using the High-Frequency Enhancer.

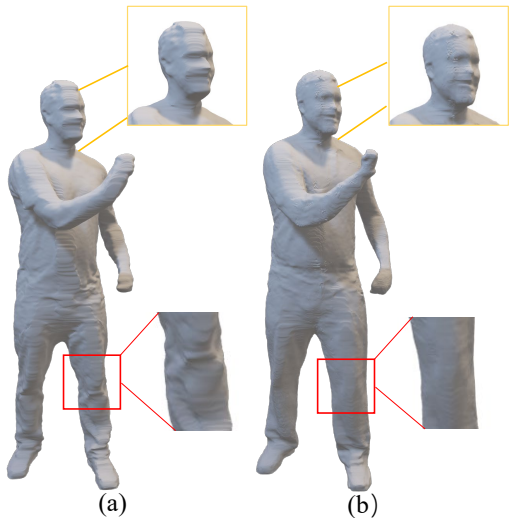

Figure 9. Ablation study on multi-view recarving strategy. With the multi-view recarving strategy (b), some artifacts can be avoided, compared with the variant without this strategy (a).

User Study. To better evaluate our method with other state-of-the-art methods, we conduct a perceptual study to ask the users about their preferences of the three methods on the following three aspects: (1) Global structure, (2) Local detail, and (3) Overall impression. In the user study, we collected 123 answers from 54 females and 69 males of different ages (6 users under 18, 113 users between 18 and 40, 2 users between 40 and 60, and 2 users above 60). The results are shown in Tab. 3, demonstrating that our approach is the most popular throughout the user study. A total of 70.73% of the users consider our methods' diversity superior to Chupa (Method B). More details about the user study can be found in the supplementary material.

Table 3. Proportion of popularity of different methods in different metrics.

Method	AG3D (A)	Chupa (B)	Ours (C)
Global Structure	18.70%	32.20%	**49.10%**
Local Detail	18.05%	29.92%	**52.03%**
Overall Impression	19.02%	29.27%	**51.71%**
Diversity of Generation	-	29.27%	**70.73%**

4.5. Ablation Study

We have conducted extensive ablation studies to validate the components and settings of our pipeline. We use the same pose guidance in quantitative comparisons. We also show and analyze different qualitative results to demonstrate the effectiveness of our modules.

Ablation on Different Pose-guidance Strategies. To validate the effectiveness of our pose-guidance, we train our model under different conditional information for pose-guidance. There are three types of conditional information in our study, SMPL [29], CSEJ (our compact spherical embedding of 3D joints), and IUV [14]. We calculate the Fid scores under the different pose conditions. As shown in Tab. 4, the CSEJ can generally improve the generation quality. While CSEJ alone is not as good as IUV on this metric, the advantage of our CSEJ is enabling precise pose control. The FID only calculates the distance between two distributions but does not measure the ability to control the pose. Fig. 7 shows generated results using different guidance under the fixed pose; our CSEJ alone achieves better pose-control results than IUV.

Table 4. Ablation study on different pose-guided strategy. We report FID scores for different combinations.

SMPL	CSEJ	IUV map	$\text{FID}_{normal} \downarrow$
✓	✗	✗	49.16
✗	✓	✗	45.73
✗	✗	✓	39.51
✗	✓	✓	**38.40**

Ablation on Multi-view Recarving Strategies and High-Frequency Enhancer. Fig. 8 demonstrates the better visual effect of using High-Frequency Enhancer. The visual comparisons presented in Fig. 9 demonstrate that the recarving strategy significantly enhances the local details of the generated human. To further measure the ability of other modules to capture local details, we conduct comparative experiments under different modules and calculate the FID scores. The results are shown in Tab. 5.

Table 5. Ablation study on multi-view recarving strategy and high-frequency enhancer. We report FID scores for different combinations.

Recarving Strategy	Enhancer	$\text{FID}_{normal} \downarrow$
✗	✗	37.39
✓	✗	31.90
✗	✓	29.51
✓	✓	**25.67**

5. Conclusion and Discussion

Conclusion. In this paper, we introduce *Joint2Human*, a novel and efficient method for directly generating detailed 3D human geometry using 2D diffusion models. We propose a compact spherical embedding of 3D joints for flexible control and the utilization of human prior. We also design a high-frequency enhancer and a multi-view recarving strategy to seamlessly integrate the details from different views into a uniform global shape, guaranteeing global structure and local details. Besides, our method can also generate high-quality 3D humans guided by text. Experimental results demonstrate that our method outperforms the state-of-the-art methods, making it ideal for advanced 3D applications.

Limitations. Although our method can produce results with various poses, supporting extreme poses is still a huge challenge, such as stooping down or standing on the head. More failure cases are shown in the supplementary material.

Broader Impact. Our method will promote the development of avatar generation, which is useful for VR/AR applications and makes up for the lack of 3D human datasets. However, this may also cause privacy and ethical problems. We suggest policymakers establish an efficient regulatory system and inform users about potential risks.

Acknowledgements. This work was supported in part by National Key R&D Program of China (2023YFC3082100), National Natural Science Foundation of China (62122058 and 62171317), and Science Fund for Distinguished Young Scholars of Tianjin (No. 22JCJQJC00040).

References

[1] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *Adv. Neural Inform. Process. Syst.*, 2022. 2, 3

[2] Naresh Bynagari. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Asian Journal of Applied Science and Engineering*, 8:25–34, 2019. 6

[3] 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. *arXiv preprint arXiv:2304.00916*, 2023. 2, 3

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 3

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[6] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 6

[7] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tuyakov, Alex Schwing, and Liangyan Gui. SDFusion: Multimodal 3d shape completion, reconstruction, and generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 3

[8] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. 2023. 3

[9] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to Generate 3D Avatars from 2D Image Collections. In *Int. Conf. Comput. Vis.*, 2023. 2, 3, 6, 7

[10] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Monocular real-time human geometry reconstruction. In *CAAI International Conference on Artificial Intelligence*, pages 594–598, 2022. 2

[11] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Fof: Learning fourier occupancy field for monocular real-time human reconstruction. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[12] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Adv. Neural Inform. Process. Syst.*, 2014. 3

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 5, 8

[15] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 5

[16] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 5

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Adv. Neural Inform. Process. Syst.*, 2020. 3

[18] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Trans. Graph.*, 2022. 2, 3

[19] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *Int. Conf. Learn. Represent.*, 2023. 2, 3

[20] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. Humangen: Generating human radiance fields with explicit priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2, 3

[21] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Trans. Graph.*, 2022. 5

[22] Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. Chupa: Carving 3d clothed humans from skinned shape priors using 2d diffusion probabilistic models. *arXiv preprint arXiv:2305.11870*, 2023. 2, 3, 5, 6, 7

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

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

[25] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Adv. Neural Inform. Process. Syst.*, 2021. 2

[26] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 3

[27] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! Text to Animatable Digital Avatars. In *International Conference on 3D Vision (3DV)*, 2024. 2

[28] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.*, 2021. 2

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 2015. 2, 3, 8

[30] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Sem-*

inal graphics: pioneering efforts that shaped the field, pages 347–353. 1998. 3

[31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Eur. Conf. Comput. Vis.*, 2020. 2

[32] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[33] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Int. Conf. Comput. Vis.*, 2021. 2

[34] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Eur. Conf. Comput. Vis.*, 2022. 2, 3

[35] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 3

[36] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[37] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2, 3

[38] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[39] Sida Peng, Chen Geng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Implicit neural representations with structured latent codes for human body modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 3

[42] Jessica Shue, Eric Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. 2022. 6

[43] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 5

[44] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2, 3

[45] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[46] Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, Shuguang Cui, and Xiaoguang Han. Get3dhuman: Lifting stylegan-human into a 3d generative model using pixel-aligned reconstruction priors. In *Int. Conf. Comput. Vis.*, 2023. 2, 3

[47] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 4

[48] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *Eur. Conf. Comput. Vis.*, 2022. 2, 3

[49] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 5

[50] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: a 3d generative model for animatable human avatars. In *Eur. Conf. Comput. Vis.*, 2022. 2, 3

[51] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2