CtrlAvatar: Controllable Avatars Generation via Disentangled Invertible Networks

Wenfeng Song[1], Yang Ding[1], Fei Hou[2,3], Shuai Li[4,5*], Aimin Hao[4], Xia Hou[1]
[1]College of Computer Science, Beijing Information Science and Technology University
[2]Key Laboratory of System Software (CAS), State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, China
[3]University of Chinese Academy of Sciences, China
[4]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[5]Zhongguancun Laboratory, China
{songwenfenga, dy1211186431}@gmail.com, houfei@ios.ac.cn, {lishuai, ham}@buaa.edu.cn, houxia@bistu.edu.cn

Abstract

As virtual experiences grow in popularity, the demand for realistic, personalized, and animatable human avatars increases. Traditional methods, relying on fixed templates, often produce costly avatars that lack expressiveness and realism. To overcome these challenges, we introduce Controllable Avatars generation via disentangled invertible networks (CtrlAvatar), a real-time framework for generating lifelike and customizable avatars. CtrlAvatar uses disentangled invertible networks to separate the deformation process into implicit body geometry and explicit texture components. This approach eliminates the need for repeated occupancy reconstruction, enabling detailed and coherent animations. The body geometry component ensures anatomical accuracy, while the texture component allows for complex, artifact-free clothing customization. This architecture ensures smooth integration between body movements and surface details. By optimizing transformations with position-varying offsets from the avatar's initial Linear Blend Skinning vertices, CtrlAvatar achieves flexible, natural deformations that adapt to various scenarios. Extensive experiments show that CtrlAvatar outperforms other methods in quality, diversity, controllability, and cost-efficiency, marking a significant advancement in avatar generation.

Code — https://github.com/1211186431/CtrlAvatar

1 Introduction

The increasing popularity of virtual experiences in gaming, social media, virtual reality (VR), and augmented reality (AR) has spurred a growing demand for realistic, personalized, and animatable human avatars. These avatars play a crucial role in enhancing user engagement and immersion, serving as the digital representation of users in various virtual environments. However, current avatar generation methods often rely on fixed templates (Loper et al. 2015) and limited control mechanisms, which fail to capture the subtle nuances of human behavior and appearance. As a result, these avatars frequently appear stiff, generic, and unrealistic, leading to a suboptimal user experience.

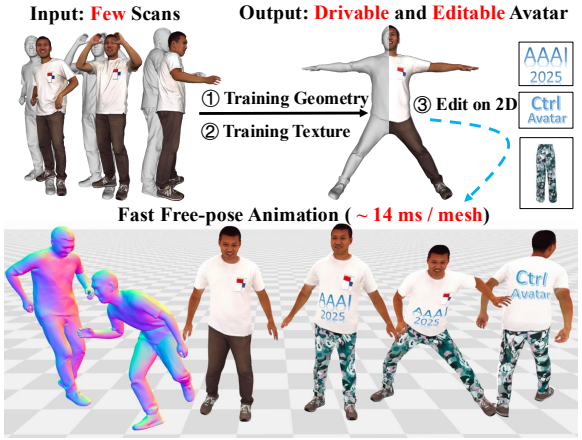

Figure 1: We present CtrlAvatar, a method for generating human avatars driven by pose parameters. CtrlAvatar features a disentangled design that allows it to achieve high-quality results with minimal training data. The model supports editable textures and enables rapid inference.

Existing methods (Huang et al. 2024; Guo et al. 2023) for human avatar reconstruction from images or videos are often limited by the need for high-quality multi-view inputs. While these approaches can generate avatars for each pose frame, they face an intrinsic challenge: **preserving the subtle nuances of human features while maintaining consistency in deformed avatars driven by poses.** This challenge arises from the need to harmonize non-rigid deformations with dynamic movements. The inherent difficulty of integrating realistic non-rigid deformations with pose-driven deformations often leads to visible artifacts, which can undermine the believability and realism of the avatars.

Inspired by the solid foundation works (Shen et al. 2023; Huang et al. 2024), the field has made significant strides in generating avatars by implicit neural networks. However, they must leverage repetitive implicit modeling to deform avatars into specific poses. The difficulty lies in the **inherent complexity of decoupling pose and texture from the underlying geometry** in a way that allows for seamless adjustments. Our CtrlAvatar introduces a new framework for creat-

*Corresponding author.

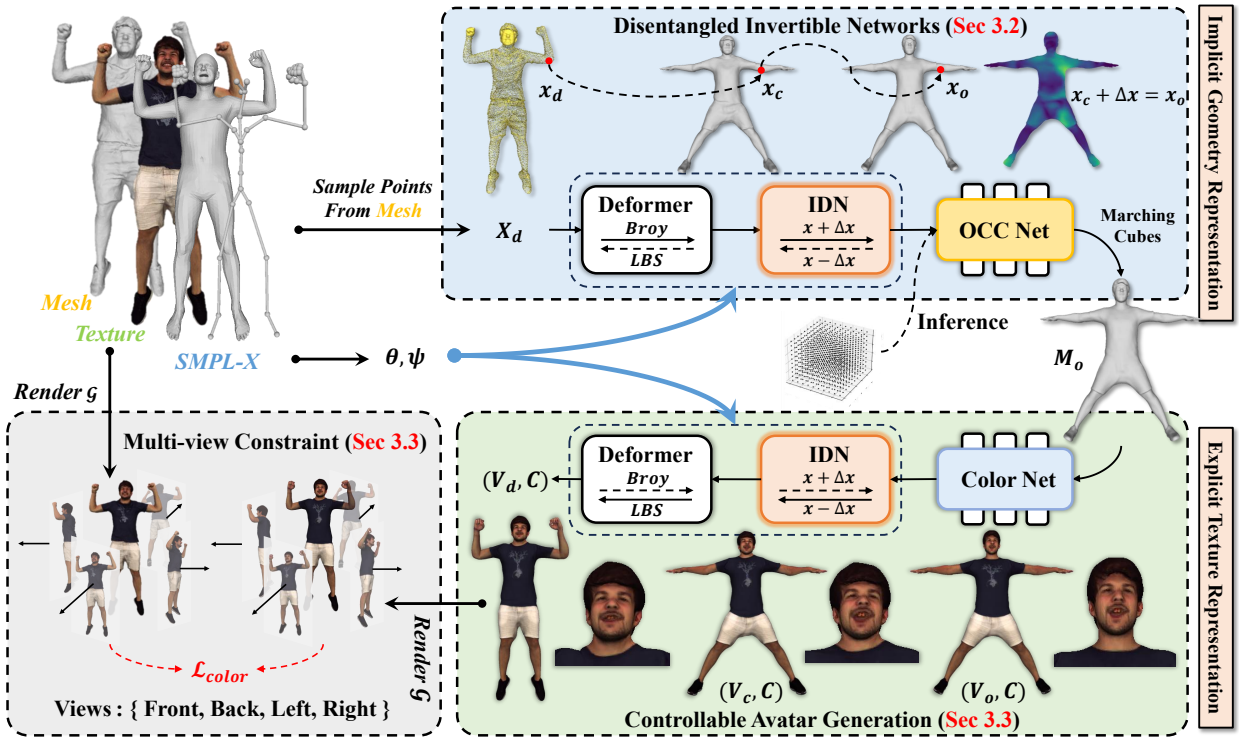

Figure 2: **Method Overview.** We propose the CtrlAvatar with two key parts: (1) Disentangled Invertible Networks, using an Invertible Delta Network to improve the avatar's implicit geometry for more realistic results; (2) Controllable Avatar Generation, employing explicit texture avatar to generate the realistic appearance.

ing expressive and customizable human avatars to overcome the limitations of current avatar generation methods. This framework includes two key components: implicit geometry representation and explicit texture representation. The former representation ensures anatomical accuracy and detailed geometry, allowing avatars to move realistically and maintain natural body structure, even during complex poses. The later representation provides advanced control over clothing and surface details, enabling users to customize styles, colors, and patterns easily, resulting in unique and diverse avatars, as shown in Fig. 1. Additionally, CtrlAvatar uses an innovative deformation neural network to enhance avatar motion, solving the stiffness problem by allowing fluid and natural movements. Meanwhile, an advanced loss function ensures the avatars maintain human-like proportions and realistic motions.

The key contributions of CtrlAvatar include:

- We introduce a **brand new CtrlAvatar**, a generation approach that separates body geometry and texture deformation, avoiding unnecessary conversions between 3D representations and ensuring smooth integration between movements and surface details.
- We present **a simple yet effective Invertible Delta Network** (IDN) that improves upon traditional Linear Blend Skinning by predicting position-varying offsets, allowing for more natural and flexible deformations.
- We develop an optimization strategy with a loss function that penalizes unrealistic deformations using multi-view 2D images, ensuring avatars maintain natural proportions and movements. Our extensive experiments demonstrate the strong performance and real-time capabilities of the CtrlAvatar framework.

2 Related Work

Human Avatar Representation. In computer graphics, capturing fine details and achieving accurate texturing in 3D avatar reconstruction remains a significant challenge. Explicit models using 3D meshes are widely used in computer graphics due to their consistent shape and structure, but they struggle with capturing clothing and fine details (Loper et al. 2015; Pavlakos et al. 2019; Saito et al. 2021). To improve this, implicit field models like signed distance fields (SDF) (Park et al. 2019; Chan et al. 2024; Qin et al. 2024) and occupancy fields (OCC) (Mescheder et al. 2019; Saito et al. 2020; Xiu et al. 2023) were introduced, which offer high-resolution details. However, their inconsistent mesh topology makes them difficult to integrate into graphic pipelines. Our approach combines the details of implicit fields with topological consistency for better integration. For texturing avatars, traditional methods (Zheng et al. 2022; Shen et al. 2023) align color fields with 3D models, but this requires extensive data to achieve high accuracy. Recent techniques like neural radiance fields (NeRF) (Mildenhall et al. 2021; Lin et al. 2024; Zhang et al. 2024) and 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023; Guédon and

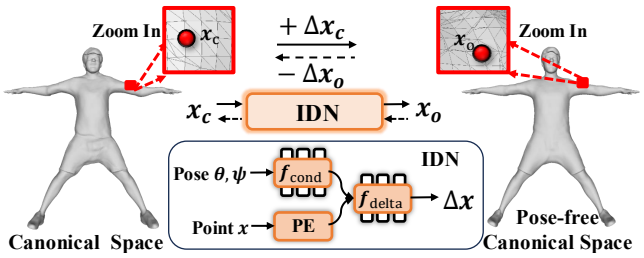

Figure 3: **Invertible Delta Network.** We leverage the IDN model to get the conditional offsets to realize non-rigid deformations (dash line denotes inevitable process).

Lepetit 2024; Qian et al. 2024) have improved rendering from multi-view videos but still struggle with full 3D model accuracy. To overcome the challenges of capturing fine details and achieving accurate texturing in 3D avatar reconstruction, we propose a method that combines the strengths of implicit geometry with explicit texture representations.

Human Avatar Deformation. Animating human avatars faces challenges with non-rigid deformations that require innovative solutions. Linear Blend Skinning (LBS) (Chen et al. 2021) is the primary technique for animating human avatars, as it more accurately replicates body deformations compared to skeletal repositioning techniques (Zhang et al. 2022). Recent improvements include CUDA acceleration (Chen et al. 2023) and enhanced control over facial expressions and hand movements with techniques like Part-Aware Sampling (Shen et al. 2023). However, LBS struggles with non-rigid deformations like clothing wrinkles. Some methods (Chen et al. 2021; Shen et al. 2023) employ conditional parameters to compute implicit fields for non-rigid deformations. Still, these require rebuilding the avatar for each pose, leading to topological inconsistencies and time-consuming mesh extraction using Marching Cubes (MC) (Lorensen and Cline 1987). Recent research (Kant et al. 2023) has explored invertible neural networks to address these shortages but hasn't yet solved them for textured avatars. Our approach overcomes these challenges by maintaining topological consistency and significantly improving animation speed, making it two orders of magnitude faster than previous works.

3 Method

Motivation and Method Overview. Our goal is to develop an efficient and parameter-driven human avatar with high-quality geometry and textures. Accurately modeling complex deformations of human avatars, particularly in dynamic scenarios, presents a significant challenge in computer graphics. Traditional approaches (Ma et al. 2020; Deng et al. 2020) often struggle to capture non-rigid deformations, such as clothing wrinkles or subtle facial expressions, resulting in unnatural and less realistic outcomes. To address this, we disentangle the deformation process into two components: geometry deformation and texture deformation. For geometry, we introduce an Invertible Delta Network (IDN) within the LBS framework to learn offsets and compute pose-free occupancy (see Sec. 3.2, top right of Fig. 2). For texture, we explicitly deform the avatars and predict colors using a color field, allowing for the creation of pose-free textures by deforming the mesh (see Sec 3.3, bottom right of Fig.2). To achieve a high-quality appearance, we render the avatar in 2D from multiple viewpoints and refine it using ground truth (GT) data (see Sec 3.3, bottom left of Fig 2). This approach results in a detailed avatar that is ready for quick and realistic animation.

3.1 Revisit of Avatar Deformation

To simulate skeletal deformation, we employ LBS to apply rigid transformations to the human avatar. In LBS, x_d represents the point within the deformation space, while x_c denotes the point in the canonical space. The linear skinning process transforms x_c into x_d. The deformation process as:

$$\begin{aligned} x_d = LBS(w, B, x_c) &= \sum_{i=1}^{55} w_i(x_c) B_i x_c, \\ \{x_c^1, \ldots, x_c^k\} &= Broy(w, B, x_d), \end{aligned} \tag{1}$$

where B_i is the rotation matrices for i-th bone in the SMPL-X (Pavlakos et al. 2019), and w_i are the weights, typically learned through a MLP. Since the canonical space points are unknown during training, we use the $Broy(\cdot)$ (Broyden 1965) to determine the solution set $\{x_c^i\}$, as done in previous work (Shen et al. 2023).

While LBS is effective for simulating skeletal deformation, it falls short in capturing non-rigid deformations, such as clothing wrinkles or facial expressions. These subtle variations are crucial for enhancing the realism and accuracy of human avatar.

3.2 Disentangled Invertible Networks

Our task in this section is to generate non-rigid deformations for human avatars while maintaining consistent topology across different poses. Previous work (Shen et al. 2023) uses the SMPL-X parameter as implicit field conditions to generate non-rigid deformations corresponding to poses. However, it introduces variability in the generated body topology, requiring new inferences for each pose, which is time-consuming. We propose Disentangled Invertible Networks to efficiently generate non-rigid deformations for various poses while maintaining consistent human body mesh topology. This network includes two key stages: Invertible Delta Network, which learns conditional offsets, and the Implicit Geometry Representation, which captures the implicit field of the human avatar.

Invertible Delta Network. We define non-rigid deformations as conditional offsets between the pose-free canonical space and canonical space. We use IDN to obtain conditional offsets, as illustrated in Fig. 3 which effectively decouples the deformations from the implicit field.

Firstly, We apply positional encoding Φ_{emb} (Niemeyer et al. 2019; Mildenhall et al. 2021) to each point x_c in the canonical space and use a MLP f_{cond} to extract conditional features:

$$\begin{aligned} \Phi_{\text{emb}} &: \mathbb{R}^3 \rightarrow \mathbb{R}^p, \\ f_{\text{cond}} &: \mathbb{R}^{|\theta|+|\psi|} \rightarrow \mathbb{R}^p, \end{aligned} \tag{2}$$

where p represents the dimension of the point after positional encoding, and θ and ψ are the parameters controlling body pose and facial expressions in SMPL-X model. Positional encoding captures high-frequency information more effectively, and mapping the conditions to the same dimension facilitates easier condition fusion.

For obtaining the conditional offset Δx_c, such as the shifting of clothing or changes in facial expressions, we fuse the point features $\Phi_{\text{emb}}(x_c)$ with the conditional features $f_{\text{cond}}(\theta, \psi)$:

$$\Delta x_c = f_{\text{delta}}\big(\Phi_{\text{emb}}(x_c) \circ f_{\text{cond}}(\theta, \psi), f_{\text{cond}}(\theta, \psi)\big), \tag{3}$$

where $\circ$ denotes the Hadamard product, which multiplies corresponding elements of the point and conditional features. The function $f_{\text{delta}}(\cdot)$ then concatenates the resulting product with the conditional features, followed by a MLP that predicts the offset. This approach ensures that when the control condition $f_{\text{cond}}(\theta, \psi) = 0$, the offset Δx_c also equals zero, thereby maintaining the integrity of the canonical space. This fusion enhances the network's ability to accurately relate conditions, such as body poses and facial expressions, to the necessary deformations.

Finally, the calculated conditional offset is added to the original point to determine the final position:

$$\begin{aligned} x_o &= IDN(x_c, \theta, \psi) = x_c + \Delta x_c, \\ x_c &= IDN^{(-1)}(x_o, \theta, \psi) = x_o - \Delta x_o, \end{aligned} \tag{4}$$

where x_o represents the final deformed point in the pose-free canonical space. By applying both the forward and inverse operations of the IDN model, we can flexibly transform between the pose-free canonical and the canonical spaces. This method effectively overcomes the limitations of traditional approaches by ensuring both flexibility and accuracy in the deformation process, resulting in more lifelike and natural avatar animations.

Implicit Geometry Representation. To effectively decouple human deformation from implicit fields, we diverge from previous methods (Chen et al. 2021; Shen et al. 2023) that rely on predicting occupancy values using canonical space points and conditions. Instead, we use x_o as the sole input for the occupancy field, where the pose-free canonical shape S_o is defined by the 0.5 level set of the occupancy field f_{occ}, representing the human surface:

$$f_{\text{occ}} : \mathbb{R}^3 \to [0, 1], \quad S_o = \{x_o \mid f_{\text{occ}}(x_o) = 0.5\}. \tag{5}$$

During network training, the deformation process of x_d is described using Equation 1 and Equation 4 as follows:

$$x_d \xrightarrow{Broy(.,B)} \{x_c^i\} \xrightarrow{IDN(.,\theta,\psi)} \{x_o^i\} \xrightarrow{Query(f_{\text{occ}}(\cdot))} x_o, \tag{6}$$

where $Query(\cdot)$ indicates applying the argmax function on the occupancy values to identify the corresponding points. By leveraging the forward process of the IDN, we obtain positions within the pose-free canonical space points, which reduces the conditional input required for the implicit field and effectively decouples condition-based deformation from the implicit field.

For training the implicit field, we adhere to the best practices outlined in (Shen et al. 2023), utilizing ground truth occupancy values and human priors as constraints to ensure precise modeling.

3.3 Controllable Avatar Generation

To achieve flexible control in avatar generation, especially when deforming avatars under SMPL-X poses, we propose an Explicit Texture Representation. This approach enables more precise deformation of both geometry and posed geometry. Furthermore, we introduce a technique called Pose-Free Texture Learning, which effectively deforms avatars while preserving texture consistency and realism.

Explicit Texture Representation. Training geometry and texture simultaneously can result in conflicts, hindering the model's ability to accurately capture both aspects (Zheng et al. 2022; Shen et al. 2023). We address this shortage by decoupling the training of geometry and texture. We start by focusing solely on training the geometry of the avatar. Once the geometry is well-learned, we use the $MC(\cdot)$ to extract a detailed and pose-free human mesh from the occupancy field:

$$M_o = (V_o, F) = MC(f_{\text{occ}}), \tag{7}$$

where V_o represents the vertices and F represents the faces of the mesh.

We obtain a clear and precise geometric representation of the avatar that is independent of any specific pose or texture. This pose-free human mesh M_o is then used as the input for the color field. This process ensures better convergence, more accurate texture mapping, and greater flexibility in avatar deformation.

Pose-Free Texture Learning. Our objective is to generate textures for human avatars that remain consistent regardless of their pose. To achieve this, we separate the deformation from the implicit field, allowing us to focus on predicting the texture of a pose-free human mesh M_o and its corresponding color field C:

$$C = f_{\text{color}}(\Phi_{\text{emb}}(V_o, N_o)), \tag{8}$$

where V_o and N_o represent the vertices and normals of the mesh M_o. Utilizing high-frequency positional encoding enhances the texture details. As the human avatar deforms, the color field C remains constant, with only the movement of the vertices being affected:

$$V_o \xrightarrow{IDN^{(-1)}(.,\theta,\psi)} V_c \xrightarrow{LBS(.,B)} V_d, \tag{9}$$

where $IDN^{(-1)}(\cdot)$, defined by Equation 4, first transforms the vertices V_o into a canonical space vertices V_c, then applies $LBS(\cdot)$ for forward skinning to achieve the final deformed space vertices V_d. Throughout this transformation, the color of the mesh vertices and faces remains unchanged, resulting in the deformed mesh being represented as $M_d = (V_d, F, C)$. Consequently, the resulting texture is also pose-free.

Multi-view Constraint. Inspired by multi-view reconstruction methods (Li et al. 2024; Hu et al. 2024; Yang et al. 2024), we render the implicit mesh M_d as multiview images to refine the 3D avatar's appearance feature. The corresponding image I_ϕ in viewpoint of M_d, can be obtained using the multi-view renderer $\mathcal{G}(\cdot)$ and the camera parameter ϕ. For

Methods	CD ↓			CD-MAX ↓			NC ↑			IOU ↑		
	ALL	Hands	Face	ALL	Hands	Face	ALL	Hands	Face	ALL	Hands	Face
HAVE-FUN (Yang et al. 2024)	6.876	**4.599**	5.797	69.649	22.883	47.654	0.879	0.772	0.882	0.945	0.759	0.904
Fast-SNARF (Chen et al. 2023)	5.227	7.561	4.149	**48.898**	39.607	21.548	**0.938**	0.785	0.932	0.936	0.604	0.887
XAvatar (Shen et al. 2023)	**5.224**	4.853	3.137	49.376	21.671	19.631	0.929	0.808	0.939	**0.963**	0.779	0.903
XAvatar-NC	5.257	4.668	3.047	48.986	20.705	18.475	0.928	0.809	0.942	0.961	**0.789**	0.908
Ours	5.327	4.651	**2.940**	49.232	**20.667**	**18.002**	0.932	**0.811**	**0.945**	0.961	0.788	**0.909**

Table 1: **Quantitative Results on SX-Humans.** We evaluate the reconstruction quality of the entire body (All) and separately assess the hands only (Hands) and face only (Face). The **bold** denotes the best performers, while the underline indicates the second best. The results verify that our approach produces high-quality human reconstructions with the most detailed facial representations.

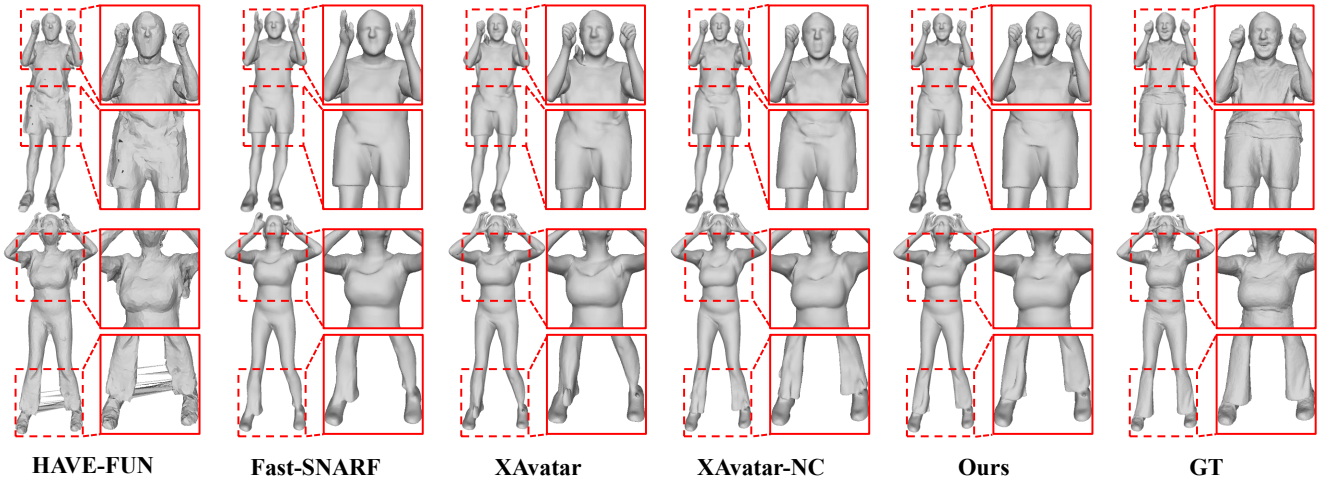

Figure 4: **Qualitative Results on the SX-Humans.** Our method avoids artifacts and captures richer details of the human body and face.

the rendered images, we apply multi-view constraints as follows:

$$\mathcal{L}_{\text{color}} = \frac{1}{4}\sum_{\phi\in\Phi}\left(\lambda_{\text{L1}}\mathcal{L}_{1,\phi} + \lambda_{\text{ssim}}\left(1 - f_{\text{ssim}}(I_\phi, I_\phi^{\text{gt}})\right) + \lambda_{\text{per}}\left\|\Psi(I_\phi) - \Psi(I_\phi^{\text{gt}})\right\|_2^2\right), \tag{10}$$

where $\Phi = \{\phi_f, \phi_b, \phi_l, \phi_r\}$ represents the four viewpoints, and $f_{\text{ssim}}(\cdot)$ stands for the Structural Similarity Index (SSIM) (Wang et al. 2004). The function $\Psi(\cdot)$ represents the VGG16 network (Simonyan and Zisserman 2014), which extracts high-level image features to improve similarity at the feature level.

During texture optimization, we freeze the weights of the LBS weights w and IDN model to ensure that the gradients updating the mesh do not affect the deformation, allowing us to focus solely on accurate color prediction.

We ensure consistency between texture and geometry using multi-view constraints, which enhances the quality of rendered images. By applying multiple 2D image-based constraints, we can generate high-fidelity textures.

4 Experiments

In this section, we present an overview of the datasets and evaluation metrics. We conduct both quantitative and qualitative comparisons of our method against the state-of-the-art in terms of reconstruction accuracy, texture quality, and inference time. Subsequently, we conduct ablation studies to confirm the efficacy of our critical design choices within key modules. Additionally, we show the result of driving and editing the human avatar. Further experimental details are provided in the Supplementary Material.

4.1 Datasets and Metrics

SX-Humans. X-Humans (Shen et al. 2023) is a comprehensive 3D clothing scanning dataset featuring textured human body scans. It includes 20 subjects, each with continuous scanning action sequences. To evaluate our method with limited data, we curated a smaller subset, SX-Humans, by selecting four scanning actions from each subject's sequences at predetermined intervals.

S-CustomHumans. CustomHumans (Ho et al. 2023) is a dataset comprising over 600 high-quality scans from a volumetric capture of 80 participants in 120 different garments

and poses. To complement the diversity of the X-Humans dataset, we curated the S-CustomHumans subset by selecting 10 subjects not present in X-Humans.

Metrics. To ensure a fair comparison of geometric accuracy, we employ Intersection over Union (IoU), Chamfer Distance (CD) measured in millimeters, and Normal Consistency (NC), following (Shen et al. 2023).

For color assessment, we render the textured human mesh using identical camera parameters and rendering techniques from four different viewpoints. The color quality is then quantified using the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) (Wang et al. 2004) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) metrics, consistent with (Weng et al. 2022) and (Yang et al. 2024).

4.2 Comparisons With State-of-the-Art Methods

We conduct comparative experiments utilizing state-of-the-art methods to generate drivable human avatars, with XAvatar (Shen et al. 2023) serving as our baseline. We enhance the XAvatar model by modifying the original code, eliminating the need to add control conditions during grid extraction. As a result, the modified XAvatar-NC approach requires only a single mesh extraction when driving. For HaveFun (Yang et al. 2024) and Editable-Humans (Ho et al. 2023), which use slightly different training inputs compared to other methods, we apply appropriate training settings to achieve the highest reconstruction quality.

Quantitative Analysis and Results. Tab. 1 presents the quantitative comparison between our CtrlAvatar and other methods on SX-Humans. The results show that our CtrlAvatar achieves the best or second-best results in nearly all evaluated metrics. Notably, our CtrlAvatar achieves the highest accuracy in the Face metric, with reductions in CD by 0.197 (6.27%) and in CD-MAX by 1.629 (8.29%) compared to the XAvatar, highlighting its effectiveness in capturing non-rigid deformations such as facial expressions.

Tab. 2 presents the quantitative texture evaluation on SX-Humans. Our CtrlAvatar excels in the LPIPS index, reducing it by 0.0812 (16.90%) compared to the baseline, which highlights its strength in capturing texture details. Although our approach scores slightly lower than others on the PSNR metric—more suited for smooth data—the visualization results clearly demonstrate superior texture quality. Tab. 3 outlines the time consumption of each module during the reconstruction and driving of the human avatar. For methods where it is not possible to calculate the time cost on individual modules, we use the overall time as a proxy. Our texture generation process takes only 8.9 ms, making it significantly faster than other methods. Additionally, due to the separation of non-rigid deformation from implicit fields in our IDN module, our CtrlAvatar does not require mesh regeneration for new poses—only deformation is needed. This results in a total processing time of just 14.33 ms, which is over **200 times faster** than XAvatar (Shen et al. 2023) and nearly 30 times faster than XAvatar-NC.

Methods	PSNR↑	SSIM↑	LPIPS↓
HAVE-FUN (Yang et al. 2024)	22.444	0.9362	0.04798
XAvatar (Shen et al. 2023)	**23.602**	0.9457	0.04805
XAvatar-NC	23.419	0.9452	0.04814
Ours	23.520	**0.9458**	**0.03993**

Table 2: **Quantitative Results of Texture on SX-Humans.** The results verify that our method performs best in texture appearance.

Methods	Ext.	Def.	Tex.	Inc.
HAVE-FUN* (Yang et al. 2024)	1022	37.5		37.5
E.H. (Ho et al. 2023)		3447		3447
Fast-SNARF (Chen et al. 2023)		3030	–	3030
XAvatar (Shen et al. 2023)	**2928**	**6.7**	417	3352
XAvatar-NC	2936	6.8	419	426
Ours	3528	14.33	**8.9**	**14.33**

Table 3: **Time Cost (ms) for Avatar Generation**. We measure the time cost in terms of extracting a mesh at 256^3 (Excluding HAVE-FUN*) resolution (Ext.), deforming mesh (Def.), generating texture (Tex.), and obtaining a new avatar pose (Inc.). The results verify that our method can achieve real-time generation, obtaining a new avatar pose. E.H. denotes the Editable-Humans (Ho et al. 2023).

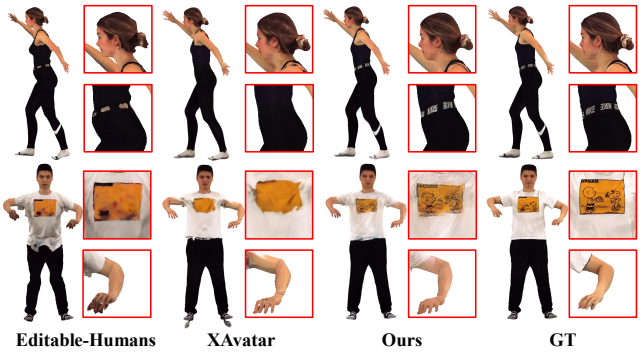

Figure 5: **Qualitative Results on S-CustomHumans with Texture.** Our method generates high-quality textures even with a limited amount of scan data for training.

Qualitative Analysis and Results. Fig. 4 provides a visual comparison between our method and others in SX-Humans. Our CtrlAvatar shows superior advantages in terms of efficiency and performance. Our CtrlAvatar requires only a single mesh extraction and effectively addresses non-rigid deformations, such as facial expressions. This allows our CtrlAvatar to achieve superior reconstruction results across varying poses. HAVE-FUN, using DMTet (Shen et al. 2021) for points extraction, has a slightly lower mesh resolution, which can make capturing fine facial details more challenging. Fast-SNARF (Chen et al. 2023) performs well overall but may struggle with accurately reproducing hand models driven by the SMPL model. XAvatar, which creates a new mesh for each pose, can sometimes show artifacts in poses

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
w/o High-frequency	30.77	0.919	0.0553 (+24.04%)
w/o Deformer	30.38	0.910	0.0543 (+21.77%)
w/o High-resolution	<u>31.80</u>	0.934	0.0507 (+13.68%)
w/ Only L1 loss	**32.08**	<u>0.942</u>	0.0491 (+9.17%)
w/ Only Two views	31.80	0.939	<u>0.0473</u> (+5.71%)
Ours	22.91	**0.942**	**0.0446**

Table 4: **Ablation Study for Texture Model.** The results verify that our texture module design achieves the best texture quality.

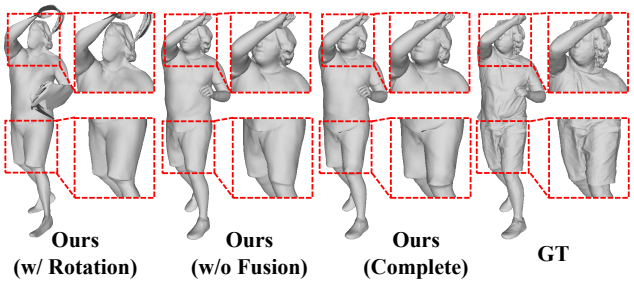

Figure 6: **Ablation Study for IDN Model.** The results verify the validity of our IDN module design, enabling the most accurate condition offsets.

that weren't well-represented in the training data. XAvatar-NC avoids the need for new meshes with each pose but may have difficulty capturing non-rigid deformations, like facial expressions, due to the absence of pose condition inputs.

We present the texture results of our CtrlAvatar and other approaches on S-CustomHumans in Fig. 5. These results demonstrate that our CtrlAvatar outperforms other approaches. By leveraging a large set of 2D images and employing the IDN model, our CtrlAvatar effectively constrains and optimizes reconstruction, resulting in superior replication of facial details and clothing textures that closely resemble real-life appearances.

While XAvatar and XAvatar-NC perform well in many respects, they may encounter challenges in reconstructing detailed textures, particularly in facial features and clothing patterns, when trained on limited data. Similarly, EditableHumans may faces difficulties in accurately reproducing clothing patterns, and geometric artifacts can further affect the overall texture quality.

4.3 Ablation Study

We conduct comprehensive ablation studies on both the IDN module and the texture module to validate the rationale behind our design choices. The visualization results of the ablation study for the IDN module are presented in Fig. 6, while Tab. 4 shows the results for the texture module.

IDN. We conduct experiments for rotation and translation to obtain the conditional offset Δx_c. The visualization results demonstrate that translation alone is sufficient to effectively learn conditional changes, while, adding rotation prevented the network from correctly training the LBS weights, as illustrated in Fig. 4. We also remove the fusion module to assess its significance. The results show that the fusion module is crucial for effectively capturing conditional information, leading to more accurate conditional offsets.

Figure 7: **Pose Driven and Editing**. We present results of editing the avatar with different clothing or patterns and driving it under various pose conditions.

Texture. For the color prediction module, we examine the effects of omitting deformation and high-frequency positional encoding. This omission resulted in significant declines in SSIM and LPIPS scores, as shown in Tab. 4. Furthermore, we experiment with different image constraint methods, such as using only L1 loss, relying on just two viewpoints, and omitting high-resolution image rendering. The experiments reveal that removing these constraints reduced the quality of the generated texture and led to a loss of detail.

4.4 Pose Driven Avatar and Avatar Editing

Thanks to the disentangled invertible networks, our CtrlAvatar system allows for flexible control over both pose and clothing texture editing. On the one hand, our approach enables the real-time generation of human animations based on arbitrary SMPL-X parameters, with consistent avatar appearance. On the other hand, it supports avatar editing through multi-view constraints, allowing for the modification of a 3D human avatar by simply adjusting 2D images (Xu et al. 2024). As illustrated in Fig. 7, our CtrlAvatar could achieve reasonable results given the conditions.

5 Conclusion

CtrlAvatar overcomes these shortages by using a disentangled invertible delta network (IDN) to separate body geometry and texture components, eliminating the need for repeated reconstruction. This approach ensures detailed, coherent animations with anatomically accurate body movements and customizable, artifact-free textures. However, our method still has limitations in reconstructing and animating loose clothing such as skirts. In future work, we plan to simplify the editing process while ensuring consistency. This includes exploring text-based editing approaches and developing tools to edit the geometric shapes of clothing directly.

Acknowledgments

This paper is supported by Beijing Natural Science Foundation (L232102), National Natural Science Foundation of China (62441201, 62272021), Beijing Science and Technology Plan Project Z231100005923039, National Key R&D Program of China (No. 2023YFF1203803), Basic Research Project of ISCAS (ISCAS-JCMS-202303), Major Research Project of ISCAS (ISCAS-ZD-202401).

References

Broyden, C. G. 1965. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19(92): 577–593.

Chan, K. Y.; Liu, F.; Lin, G.; Foo, C. S.; and Lin, W. 2024. Fine Structure-Aware Sampling: A New Sampling Training Scheme for Pixel-Aligned Implicit Models in Single-View Human Reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 964–971.

Chen, X.; Jiang, T.; Song, J.; Rietmann, M.; Geiger, A.; Black, M. J.; and Hilliges, O. 2023. Fast-SNARF: A fast deformer for articulated neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 11796–11809.

Chen, X.; Zheng, Y.; Black, M. J.; Hilliges, O.; and Geiger, A. 2021. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11594–11604.

Deng, B.; Lewis, J. P.; Jeruzalski, T.; Pons-Moll, G.; Hinton, G.; Norouzi, M.; and Tagliasacchi, A. 2020. Nasa neural articulated shape approximation. In *Proceedings of the European Conference on Computer Vision*, 612–628.

Guédon, A.; and Lepetit, V. 2024. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5354–5363.

Guo, C.; Jiang, T.; Chen, X.; Song, J.; and Hilliges, O. 2023. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12858–12868.

Ho, H.-I.; Xue, L.; Song, J.; and Hilliges, O. 2023. Learning locally editable virtual humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21024–21035.

Hu, L.; Zhang, H.; Zhang, Y.; Zhou, B.; Liu, B.; Zhang, S.; and Nie, L. 2024. 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*, 634–644.

Huang, Y.; Yi, H.; Xiu, Y.; Liao, T.; Tang, J.; Cai, D.; and Thies, J. 2024. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision*.

Kant, Y.; Siarohin, A.; Guler, R. A.; Chai, M.; Ren, J.; Tulyakov, S.; and Gilitschenski, I. 2023. Invertible neural skinning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8715–8725.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Li, Z.; Zheng, Z.; Wang, L.; and Liu, Y. 2024. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19711–19722.

Lin, W.; Zheng, C.; Yong, J.-H.; and Xu, F. 2024. Relightable and animatable neural avatars from videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3486–3494.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics*, 34(6): 248:1–248:16.

Lorensen, W. E.; and Cline, H. E. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21(4): 163–169.

Ma, Q.; Yang, J.; Ranjan, A.; Pujades, S.; Pons-Moll, G.; Tang, S.; and Black, M. J. 2020. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6469–6478.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4460–4470.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Niemeyer, M.; Mescheder, L.; Oechsle, M.; and Geiger, A. 2019. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5379–5389.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 165–174.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10975–10985.

Qian, Z.; Wang, S.; Mihajlovic, M.; Geiger, A.; and Tang, S. 2024. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5020–5030.

Qin, M.; Liu, Y.; Xu, Y.; Zhao, X.; Liu, Y.; and Wang, H. 2024. High-Fidelity 3D Head Avatars Reconstruction through Spatially-Varying Expression Conditioned Neural Radiance Field. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4569–4577.

Saito, S.; Simon, T.; Saragih, J.; and Joo, H. 2020. 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*, 84–93.

Saito, S.; Yang, J.; Ma, Q.; and Black, M. J. 2021. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2886–2897.

Shen, K.; Guo, C.; Kaufmann, M.; Zarate, J. J.; Valentin, J.; Song, J.; and Hilliges, O. 2023. X-avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16911–16921.

Shen, T.; Gao, J.; Yin, K.; Liu, M.-Y.; and Fidler, S. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34: 6087–6101.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Weng, C.-Y.; Curless, B.; Srinivasan, P. P.; Barron, J. T.; and Kemelmacher-Shlizerman, I. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16210–16220.

Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; and Black, M. J. 2023. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 512–523.

Xu, Y.; Gu, T.; Chen, W.; and Chen, C. 2024. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*.

Yang, X.; Chen, X.; Gao, D.; Wang, S.; Han, X.; and Wang, B. 2024. Have-fun: Human avatar reconstruction from few-shot unconstrained images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 742–752.

Zhang, H.; Chen, B.; Yang, H.; Qu, L.; Wang, X.; Chen, L.; Long, C.; Zhu, F.; Du, D.; and Zheng, M. 2024. Avatarverse: High-quality & stable 3d avatar creation from text and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7124–7132.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.

Zheng, Y.; Abrevaya, V. F.; Bühler, M. C.; Chen, X.; Black, M. J.; and Hilliges, O. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13545–13555.