GBCN:3D-Guided Brownian Bridge Diffusion for Clothed Human Reconstruction via Normal Integration

Anonymous ICCV submission

Paper ID 8220

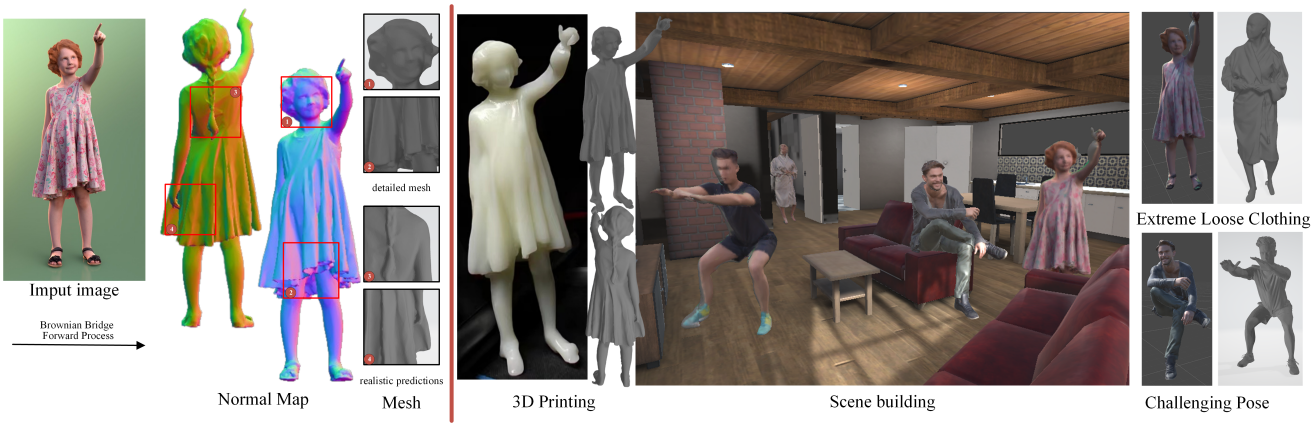

Figure 1. GBCN is a novel human body reconstruction network based on a diffusion model. Given an image, GBCN utilizes the Brownian Bridge diffusion process to map the original image to 3D features in the latent space. It refines normal maps to reconstruct a detailed clothed human mesh. By employing a consistency loss function and an SMPL-X prior decoupling module, GBCN achieves an effective extension of the Brownian Bridge diffusion process from 2D to 3D, bridging large domain gaps. Our method demonstrates excellent performance on in-the-wild images and shows significant potential in real-world applications such as 3D printing and scene construction.

Abstract

Creating high-quality 3D models of clothed humans from a single image is crucial for real-world applications. Existing models often make insufficient assumptions about the geometric details of invisible areas, leading to less refined outputs. Despite recent progress, 3D diffusion models still treat diffusion as a conditional generation process, significantly impacted by domain gaps. We addressed this by proposing GBCN (3D-Guided Brownian Bridge Diffusion for Clothed Human Reconstruction via Normal Integration) as a novel solution. Specifically, we utilized the Brownian Bridge diffusion process to directly map the image domain to the spatial domain through bidirectional diffusion, avoiding conditional information leverage. Additionally, we progressively guided the diffusion process with 3D pose priors to bridge the gap between 2D and 3D domains. To better learn fine 3D geometric details in the latent space, we employed a consistency loss function to enhance the model's robustness while maintaining sensitivity to fine geometric details, ultimately regressing detailed front and back normal maps. Finally, we used these detailed normal maps to reconstruct a fine 3D clothed model through bidirectional normal integration and Poisson surface reconstruction. Our method significantly improves geometric details, including invisible areas, and shows robustness under various poses and loose clothing. Experiments demonstrate that due to fine normal reconstruction, GBCN exhibits impressive performance in geometric reconstruction pipelines, showing enhanced robustness in complex scenarios. We also applied our method to virtual scene construction and 3D printing, showcasing its broad potential for practical applications.

1. Introduction

High-quality 3D human avatar models have crucial applications in fields such as virtual reality, education, and gaming. Creating a detailed clothed human model traditionally requires experienced artists and expensive scanning equipment [19, 53]. Generally, single RGB images of humans

are more accessible. Nevertheless, reconstructing clothed humans from a single RGB image still faces three key technical challenges: **(i)** Single RGB images suffer from severe self-occlusion, making inference about large invisible areas a major challenge. **(ii)** In practical applications, people in images may pose with large angles or wear extremely loose clothing, demanding high robustness from the model. **(iii)** Images may contain many details such as clothing wrinkles or accessories, requiring the model to accurately reconstruct these high-frequency details.

Existing models [5, 9, 31, 65, 66, 76, 77, 83, 89–91] typically use various methods to map 2D images to 3D space. Most methods [76, 77, 83, 91] focus on improving the transition from 2D features to 3D using SMPL [59] guidance, but they neglect 3D information during 2D feature extraction. Some methods [89] use fixed learnable embeddings for 3D feature learning. SIFU [90] improves on these methods by using SMPL priors through an attention mechanism to learn 3D information from images and predict the overall mesh. Despite success in handling poses and loose clothing, these methods lack exploration of surface details, limiting the generation of high-precision geometric models, as shown in Figure 2. Therefore, a new approach is needed that maintains robustness to clothing and poses while achieving good performance in geometric details, including invisible areas.

Diffusion models [24, 57, 60, 64, 70, 70, 74], compared to traditional GAN-based models, exhibit better stability and controllability [11, 12]. These methods improve the quality of generated samples by avoiding information loss through a gradual diffusion process from the source domain to the target domain. The Brownian Bridge Diffusion Model (BBDM) [44] models image-to-image translation [32] as a stochastic Brownian Bridge process, achieving progress by directly mapping between two image domains. SiTH [23] successfully applied diffusion models to the field of 3D human reconstruction and achieved impressive results. SiTH first hallucinates back-view appearances through an image-conditioned diffusion model, followed by the reconstruction of full-body textured meshes using both the front and back-view images. However, in essence, the diffusion module in SiTH is still essentially a 2D image-to-image conversion task. While recent diffusion models show promise in image synthesis, their application to 3D reconstruction [23, 30, 38], faces fundamental limitations: **(i)** As BBDM states, most diffusion models treat diffusion as conditional generation and lack direct mapping methods [44]. **(ii)** Unlike image translation tasks, transitioning from 2D images to 3D space often faces domain gaps, making effective mapping difficult without prior conditions. **(iii)** Some text-guided conditional diffusion models also require more computation.

Our key insight is that bidirectional diffusion between two domains, rather than conditional generation, more effectively guides the model in learning the mapping process. This motivates us to extend BBDM to 3D reconstruction through a prior decoupling module to address significant domain gaps. Specifically, we perform bidirectional diffusion using the original image and normal maps in the model and progressively refine geometric details in the latent space. To better match the mapping between the two domains, we gradually introduce SMPL-X [59] priors into the diffusion process using a cross-attention mechanism and use a consistency loss to guide the model in maintaining sensitivity to geometric details. Through this approach, we restored detailed front and back normal maps from the latent space and used normal integration and Poisson reconstruction [77] to generate the corresponding 2.5D surfaces. We then used SMPL-X to merge the human body mesh, achieving detailed reconstruction of the clothed human mesh without additional parameter training. Experiments show that our model benefits significantly from 3D information, proving it is not just a 2D image conversion task, as detailed in our experimental section.

Through extensive experiments, GBCN has demonstrated superior geometric reconstruction quality. It surpasses existing state-of-the-art (SOTA) methods. It shows excellent robustness in handling complex poses and loose clothing. GBCN excels in preserving detailed geometric features on surfaces, including invisible areas. We have extended the reconstruction results to practical applications. These include 3D printing and scene construction, showcasing its wide applicability. Our main contributions include:

- We improved the Brownian Bridge diffusion process and extended it to human reconstruction, establishing a mapping from 2D images to the 3D latent space and achieving richer reconstruction details.
- The normal map prediction module in GBCN is both simple and portable. Applied to recent works, it showed varying degrees of performance improvement.
- The GBCN pipeline achieved significant progress, surpassing previous methods across all metrics. Its strong performance on real-world images enabled applications in 3D printing and scene construction, highlighting its potential for state-of-the-art (SOTA) reconstruction.

2. Related Work

Diffusion models. Recent advances in diffusion models [10, 24, 64, 70] have demonstrated superior performance in high-fidelity image synthesis, particularly in preserving geometric details and handling multi-modal distributions. However, their application to 3D human reconstruction poses unique challenges. Most generative models [28, 61, 69, 73, 79] focus on object generation in the 3D creation process but are limited in generating models with realistic details due to the difficulty in collecting large-

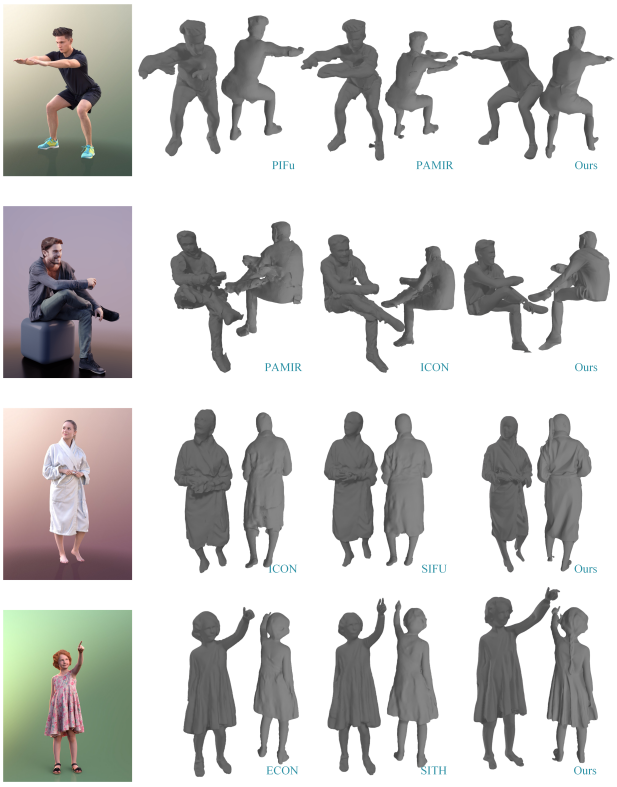

Figure 2. Comparison between GBCN and State-of-the-Art (SOTA) Methods for 3D Human Inference on In-the-Wild Images. Existing SOTA methods consistently underperform when handling complex poses or loose clothing, often producing inference results that lack sufficient detail. In contrast, GBCN effectively overcomes these challenges, yielding high-quality, highly detailed results.We use slightly different angles to fairly show the shortcomings of each model.

scale ground truth. Recent works [8, 23, 27, 30, 38, 67, 80] still treat diffusion models as conditional generative models, which lack direct mapping methods. Some methods integrate image-to-text models [8, 30] or directly use text [38, 51, 52, 55, 71, 80, 90] as conditional input in the U-Net during the reverse process to guide diffusion towards the target domain, which introduces high computational costs. BBDM [44] is a novel diffusion model that first models the image-to-image translation process as a stochastic Brownian Bridge process, making progress by directly mapping between two image domains. Our method extends the Brownian Bridge diffusion to the 3D domain, addressing the significant domain gap and achieving good performance.

Implicit-function-based Reconstruction. Implicit representations, such as occupancy fields and SDF, effectively describe 3D clothed humans in various scenarios, including challenging poses and loose clothing. Recent methods [1, 20, 65, 66] employ neural implicit functions to directly regress 3D surfaces from single images by learning a continuous mapping from 2D features to occupancy/SDF values, achieving topology-aware reconstruction. To address the inherent ambiguity in single-view reconstruction, several methods [6, 7, 9, 21, 22, 30, 31, 50, 76, 77, 83, 89, 91] incorporate SMPL [59] priors to provide strong geometric constraints and anatomical knowledge, thereby improving the accuracy of both 2D feature extraction and 3D surface prediction. For instance, GTA [89] uses a transformer with fixed learnable embeddings to convert image features into 3D planar features. SIFU [90] goes further by incorporating SMPL-X priors into 2D image extraction and using a transformer to extract multi-view features and regress the complete mesh. Despite progress, these methods lack fine surface details and fail to assume reasonably about invisible surfaces.

Explicit-shape-based Reconstruction. In contrast to implicit methods that learn continuous functions, explicit shape-based approaches [14, 15, 37, 39–42, 45–48, 68, 86, 87] directly estimate discrete representations (e.g., meshes [36, 54, 58, 59, 78], depth maps [85], or point clouds [85]) and utilize clothing offsets [2–4, 43, 75] to model surface details. While these methods offer direct geometric control, their fixed topology assumption fundamentally limits their ability to handle significant topology changes in loose clothing, model complex self-intersections, and represent fine-scale surface details. ECON [77] constructs partial surfaces by integrating front and back normal maps and stitching them with the implicit IF-Net and SMPL meshes using Poisson surface reconstruction. Although ECON's hybrid strategy enhances surface reconstruction, the use of less detailed normal maps and the polishing process of IF-Net can result in meshes lacking detail.

NeRF-based Reconstruction. Neural Radiance Fields (NeRF) [17, 18, 25, 34, 35, 49, 56, 74, 88] revolutionize novel view synthesis by learning continuous volumetric representations through implicit MLPs and differentiable volume rendering. This enables high-fidelity reconstruction of both geometry and appearance. SHER [26] and ELICIT [29] aim to reconstruct 3D clothed humans from a single image. However, applying NeRF to single-image human reconstruction faces several fundamental challenges: computational complexity due to volumetric optimization requiring extensive resources, view dependency causing significant quality degradation with limited input views, geometry ambiguity resulting in imprecise recovered geometry in occluded regions without multi-view constraints, and training efficiency issues with per-scene optimization making it difficult to learn generalizable priors.

Inspired by previous work, our method achieved impressive results in 3D clothed human reconstruction across various scenarios. We used the Brownian Bridge diffusion process, employing bidirectional diffusion with the original im-

age and ground-truth normal maps to guide the network in refining geometric details in the latent space. By progressively incorporating SMPL-X priors into the diffusion process, we successfully addressed large domain gaps. We also applied a combined consistency loss to constrain the model, ensuring robustness while maintaining sensitivity to fine geometric details.

3. Method

Our pipline consists of three key components: **(i)** A Brownian Bridge diffusion framework that learns the mapping between 2D and 3D spaces(Section 3.1). **(ii)** A novel consistency loss that ensures geometric detail preservation (Section 3.2). **(iii)** A reconstruction pipeline that leverages normal maps for high-fidelity surface generation (Section 3.3). Specifically, GBCN first uses the Brownian Bridge diffusion process to learn 3D information in the latent space through bidirectional diffusion from the original image and normal maps, while attempting to refine these details. To address large domain gaps, we use SMPL-X priors and progressively integrate them into the diffusion process. Once the model learns the mapping relationship during diffusion, we can directly regress detailed front and back normal maps from the latent space without intermediate steps. To enhance the model's sensitivity to fine details while maintaining overall perception, we apply a consistency loss in the latent space to constrain the network and guide it in refining geometric details. Finally, we use the regressed detailed normal maps to reconstruct the front and back 2.5D surfaces, completing them with human priors. For the complete pipeline, refer to Fig.3.

3.1. 3D-Guided Brownian Bridge Diffusion.

Given an RGB image input I and its corresponding normal maps N, we formulate the reconstruction process as a bidirectional mapping problem. Unlike traditional diffusion models that treat the endpoint as noise, we innovatively use the clean input image x_t as the endpoint xt, while setting the target normal maps as the starting point x_0. This design allows us to: **(i)** Maintain high-fidelity image features throughout the diffusion process. **(ii)** Explicitly learn the geometric mapping in the latent space. **(iii)** Better preserve structural consistency during diffusion. To accelerate training and inference, we retrained the popular VQGAN and performed diffusion in the latent space. For details on traditional diffusion models, please refer to the supplementary materials. We represent the state distribution at each time step of a Brownian bridge process starting from point $(\boldsymbol{x}_0) \sim q_{data}(\boldsymbol{x}_0)$ at $t = 0$ and ending at point x_t at $t = T$ can be formulated as:

$$p(x_t \mid x_0, x_T) = \mathcal{N}((1 - \frac{t}{T})\boldsymbol{x}_0 + \frac{t}{T}\boldsymbol{x}_T, \frac{t(T-t)}{T}\boldsymbol{I}) \quad (1)$$

The two ends of the process are anchored by x_0 and x_t, forming a bridge spanning two domains.

When predicting the front normal maps, we applied the diffusion process to the original image pairs and made progress. However, for the back normal maps and original image pairs, the model failed to estimate the mapping between the two domains due to the large domain gap (see experiment section for details). Although back normal maps differ in viewpoint, they maintain several critical correlations with the input image: **(i) Geometric consistency:** The overall body structure remains coherent. **(ii) Feature correspondence:** Local geometric details (e.g., clothing wrinkles) follow consistent patterns. **(iii) Topological relationships:** The spatial arrangement of body parts preserves relative positioning. We leverage these inherent correlations through a cross-attention mechanism to guide the generation of occluded regions. The specific formula is as follows:

$$CrossAttn(z_i, h) = Softmax(\frac{(W^Q h)(W^K z_i)^T)}{\sqrt{d}})(W^V z_i) \quad (2)$$

where $\boldsymbol{h}$ represents the embedding extracted from the SMPL-X priors via a feature encoder. $\boldsymbol{z}_i$ is the intermediate feature sampled during the diffusion process, and W^Q, W^K, and W^V are learnable parameters. We believe this design allows the model to selectively integrate prior knowledge and adaptively handle viewpoint changes while maintaining geometric consistency. We follow the traditional cross-attention design, to add norm layers and residual connections after each layer. Our feature extractor is multi-scale and performs cross-attention operations at different stages of the diffusion process. This design enables the model to effectively learn spatial information, including back-view perspectives. It also allows the model to decouple features from the original image, successfully mapping front-view and back-view 3D information.

3.2. Normal consistency loss.

The key challenge in normal map prediction lies in preserving geometric accuracy while maintaining spatial consistency. Unlike regular RGB images with purely visual color variations, normal maps encode crucial geometric information through their RGB channels, specifically representing surface orientations in tangent space. This special texture pattern makes the lighting effects resemble those of real geometric shapes. By learning this pattern, the model can refine spatial details in the 3D latent space and produce logical, detailed textures, avoiding fragmented reconstruction errors. An intuitive method is to constrain the model's learning of texture patterns through color consistency loss, but this faces challenges in bidirectional diffusion latent space. Despite perceptual consistency, VQGAN-encoded high-dimensional features statistically deviate from the original input, leading to uncer-

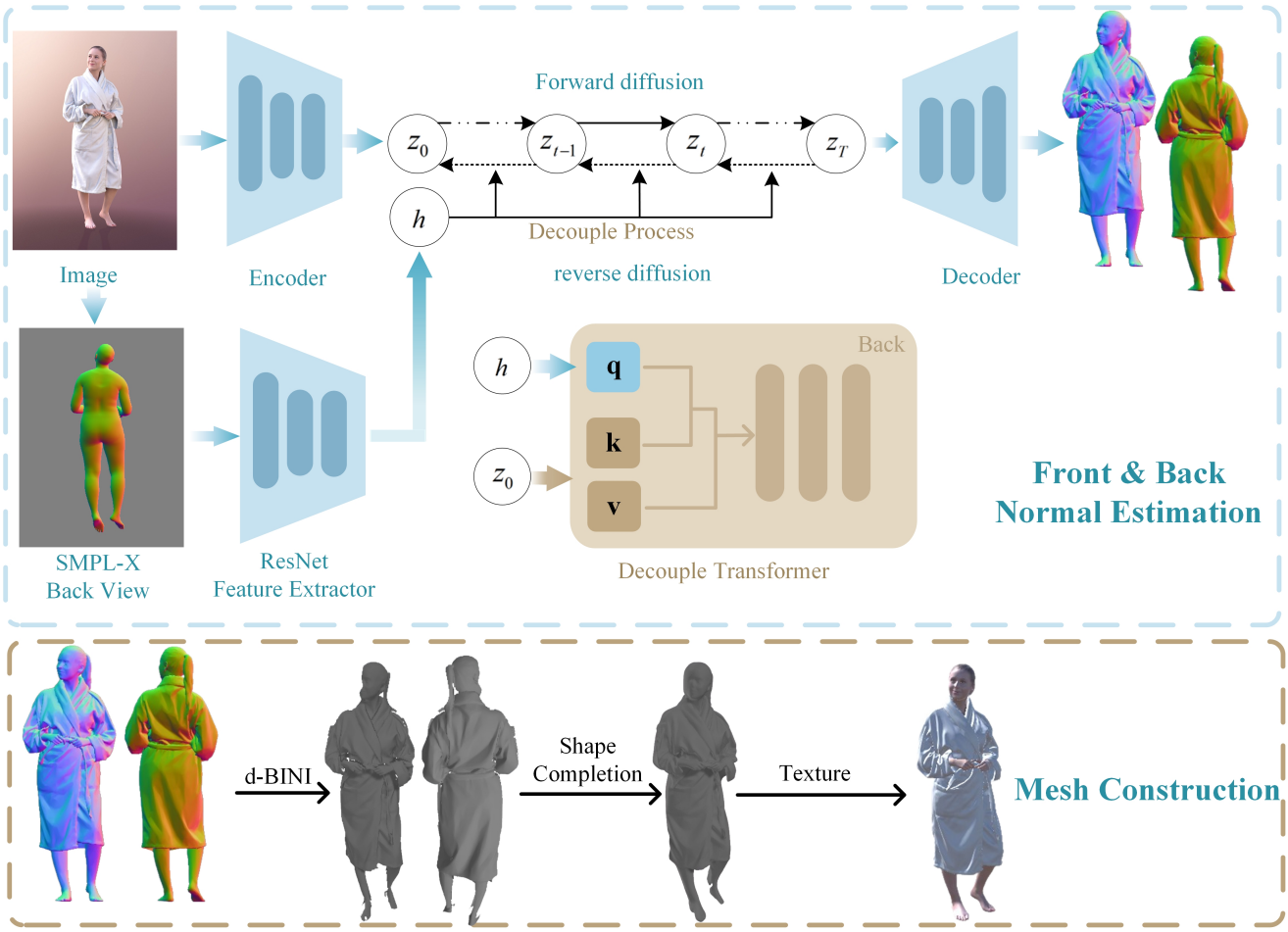

Figure 3. Given an image, GBCN utilizes the Brownian bridge diffusion process (§3.1) to directly predict the front and back normal maps of a clothed human body. During the back normal map prediction, a cross-attention mechanism is employed to integrate SMPL back-view priors, where a prior-mixing strategy incorporates these priors as query inputs, thereby enhancing the model's capability to infer occluded regions from the image. To facilitate more accurate learning of normal map color patterns, a series of hybrid consistency losses (§3.2) are applied to guide the network's training process. The subsequent step involves 3D surface reconstruction (§3.3): specifically, under the guidance of SMPL-X estimates, GBCN converts the normal maps into incomplete 2.5D front and back surfaces and reconstructs the missing geometric details. Finally, a pre-trained texture prediction network is used to produce a high-quality textured mesh that can be directly employed in practical applications.

tainties in conventional losses within the latent space. Our method uses cosine similarity loss in the latent space to constrain the direction of high-dimensional vectors. We have the following equation:

$$\mathcal{L}_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \cos_{\text{sim}}(\mathbf{a}, \mathbf{b}) \tag{3}$$

where $\mathbf{a}$ and $\mathbf{b}$ represent the predicted vector and the ground truth vector, respectively. $\cos_{\text{sim}}(\mathbf{a}, \mathbf{b})$ is the cosine similarity of vectors $\mathbf{a}$ and $\mathbf{b}$, defined as:

$$\cos_{\text{sim}}(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}} \tag{4}$$

To capture both local and global geometric patterns, we were inspired by the traditional Discrete Fourier Transform (DFT) [16, 33, 62, 81, 82, 92] to convert high-dimensional features in the latent space into frequency domain components to describe vector patterns. We perform the Discrete Fourier Transform on the original latent space vectors using a two-dimensional DFT:

$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \cdot e^{-i2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{5}$$

where $f(x, y)$ represents the pixel value at coordinate (x, y), and $F(u, v)$ represents the complex frequency value

Figure 4. Qualitative results on in-the-wild images. We present nine examples of 3D humans with detailed clothing reconstructed from wild images. These examples include loose clothing and challenging poses. For each example, we show the input image, the predicted front and back normal maps, and the reconstruction results. Our method shows impressive results.

at coordinate (u, v) in the frequency domain. The amplitude and phase of the frequency are given by:

$$\begin{cases} \mid F(u,v) \mid = \sqrt{R(u,v)^2 + I(u,v)^2}, \\ \angle F(u,v) = \arctan\left(\dfrac{I(u,v)}{R(u,v)}\right) \end{cases} \tag{6}$$

where $R(u, v)$ and $I(u, v)$ are the real and imaginary components of $F(u, v)$. To compute the frequency distance between the real and generated images, we use the Euclidean distance:

$$\mathcal{L}_{ffl} = d\left(Fr, Ff\right) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{u=0}^{N-1} \left| Fr(u,v) - Ff\left(u,v\right) \right|^2 \tag{7}$$

In addition to the aforementioned losses, our research found that MSE loss corrects larger discrepancies but neglects some details. This results in many reconstruction artifacts and poor performance during the reconstruction process. Specific experimental results are discussed in the experiment section. Therefore, we only used L1 loss for constraints.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{11} + \lambda_2 \mathcal{L}_{ffl} + \lambda_3 \mathcal{L}_{\cos} \tag{8}$$

where λ_1,λ_2,λ_3 are the weights attributed to each loss.

3.3. 3D Surface Reconstruction

For the detailed normal maps, we use variational normal integration for explicit reconstruction of the front and back surfaces. Specifically, we apply customized bilateral normal integration to form 2.5D surfaces and integrate them using Poisson reconstruction and SMPL-X surfaces [77]. Previous methods added an implicit function to polish and complete the surfaces to address artifacts caused by various poses or incomplete photos. However, this often made the predicted surfaces too smooth. Our method uses the detailed and complete normal maps. This allows us to use Poisson reconstruction to directly reconstruct the 2.5D surfaces without implicit polishing, achieving good results.

4. Experiments

4.1. Dataset and Metrics

We trained our model on Thuman2.0 and tested it on the CAPE dataset. The 526 human scans in Thuman2.0 were divided into 490 for training, 15 for We trained our model on Thuman2.0 and tested it on the CAPE dataset. The 526 human scans in Thuman2.0 were divided into 490 for training, 15 for validation, and 21 for testing. Ground-truth

Method	Publication	CAPE-NFP			CAPE-FP			THuman2.0		
		Chamfer ↓	P2S ↓	Normal ↑	Chamfer ↓	P2S ↓	Normal ↑	Chamfer ↓	P2S ↓	Normal ↑
w/o SMPL-X body prior										
PIFu[65]	ICCV 2019	3.2551	2.5492	0.7651	1.8391	1.7612	0.8596	1.2097	1.1324	0.7707
PIFuHD[66]	CVPR 2020	2.9766	2.3700	0.7680	1.5228	1.4860	0.8741	0.9953	0.9676	0.7912
PaMIR[91]	TPAMI 2021	7.1599	3.3852	0.6372	6.0132	3.2895	0.6759	1.0898	1.0169	0.7965
SITH[23]	CVPR 2024	2.8753	2.1244	0.7822	2.1158	1.6769	0.8353	0.9677	0.9052	0.7857
GBCN-PIFuHD	-	2.6909	2.1533	0.8014	1.2622	1.2803	0.9035	0.7485	0.6627	0.8149
GBCN-SITH	-	2.6617	1.9239	0.8164	1.8256	1.4044	0.8533	0.8098	0.8041	0.8113
w/ SMPL-X body prior										
ICON[76]	CVPR 2022	1.5521	1.1997	0.8584	0.9962	0.8881	0.9209	0.6173	0.5956	0.8518
ECON[77]	CVPR 2023	1.8541	1.5721	0.8414	1.1779	1.1369	0.8989	0.6741	0.6349	0.8377
D-IF[83]	ICCV 2023	1.8324	1.5637	0.8634	1.1493	1.1435	0.8893	0.6725	0.6331	0.8364
GTA[89]	NeurIPS 2023	1.8866	1.4929	0.8278	1.1496	0.9937	0.9021	0.5809	0.5601	0.8412
SIFU[90]	CVPR 2024	1.5758	1.2803	0.8545	1.0549	0.9702	0.9040	0.5774	0.5595	0.8512
GBCN-ICON	-	1.2654	1.0980	0.8860	0.8424	0.8251	0.9277	0.5844	0.5670	0.8497
GBCN(Ours)	-	**1.2189**	**0.9793**	**0.9011**	**0.7935**	**0.7083**	**0.9337**	**0.5654**	**0.5476**	**0.8702**

Table 1. Quantitative evaluation against SOTA (Section 4.1). All models use a resolution of 256 for marching cubes and ground-truth SMPL-X models are used during testing.

SMPL-X models were used during training, and PIXIE was employed for inference. To enhance robustness, we added random noise to the original images during training to simulate real environment occlusions. All training and testing were conducted on an NVIDIA GeForce RTX 3090 GPU. We adhered to the common design of previous approaches, using Chamfer and Point-to-Surface (P2S) distances to compute larger errors and measuring normal errors to assess reconstruction quality.

4.2. Evaluation

We compared GBCN with state-of-the-art (SOTA) models across varying data volumes. To evaluate our model, we split CPAE into CAPE-FP and CAPE-NFP subsets following established conventions. Ground-truth SMPL models were provided for methods requiring SMPL-X body priors. The results are presented in Tab.1. To further validate our approach, we applied GBCN to other frameworks, including PIFuHD, ICON, and SiTH. By replacing their normal prediction modules, we observed improved performance metrics despite constraints from other internal modules. GBCN effectively mapped input images to 3D space and generated detailed normal maps. Qualitative results are displayed in Fig.4. Tests on in-the-wild data confirmed that our model produced accurate normal maps and reconstructed detailed clothed human meshes, performing particularly well with loose clothing and complex poses.

4.3. Ablation Studies

The quantitative results of the ablation study are summarized in Table 2. Without the SMPL prior, even the SOTA image-to-image translation model BBDM shows only limited improvement. Incorporating the SMPL prior through our decoupling module significantly enhances performance, highlighting the importance of 3D information in generating normal maps, which cannot be treated as a simple 2D transformation task. Qualitative results are presented in Fig.5. Without the decoupling module and human pose priors, the network struggles to infer detailed back normal maps from front images. The large domain gap between back normal maps and original images hinders effective mapping to the 3D latent space. Integrating SMPL-X priors into the diffusion model's information flow improves the network's understanding of back pose structures and prevents distortions.

Method	Chamfer ↓	P2S ↓	Normal ↑
ECON	0.6741	0.6349	0.8377
GBCN w/o SMPL + L1	0.6332	0.6054	0.8448
GBCN w/ L2	0.6354	0.6048	0.8478
GBCN w/ L1	0.6046	0.5995	0.8574
GBCN w/ L1 + FFL	0.5725	0.5537	0.8633
GBCN w/ L1 + Cos	0.5833	0.5638	0.8637
GBCN(ours)	**0.5654**	**0.5476**	**0.8702**

Table 2. Ablation experiments of GBCN on Thuman2.0 dataset. The experiments showed that each module and loss in our method improved the model's performance.

During network training, we used different loss mixing strategies to supervise the network and compared the cross-entropy (MSE) loss values between the predicted and ground-truth normal maps, as shown in Tab.2 and Fig.6. Our qualitative results show that while L2 supervision slightly reduced loss values, it caused large artifacts in the images, leading to instability in subsequent surface reconstruction. Considering the impact of different losses, we adopted the mixed loss strategy shown in Equation 8.

Figure 5. We conducted ablation experiments to evaluate the conditional decoupling module by selecting a subset of images from the training set and feeding them into the converged model. Qualitative results indicate that, without the SMPL-X prior, the network is unable to effectively learn the normal map for unseen regions.

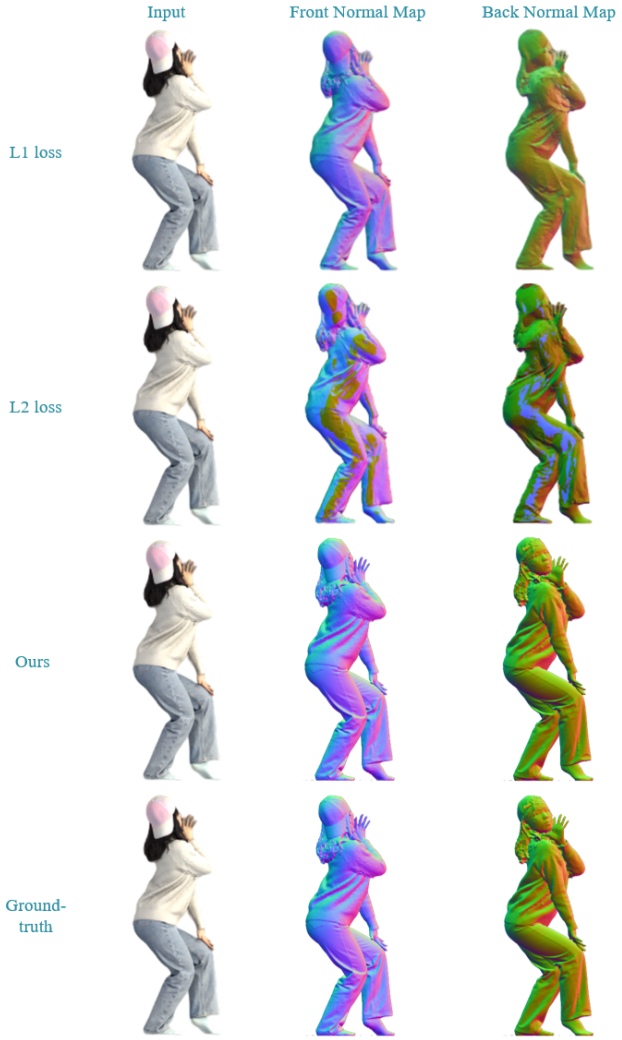

Figure 6. Qualitative evaluation of different loss combinations demonstrates the effectiveness of our chosen combined consistency loss. Notably, although the L2 loss achieved lower reconstruction error in quantitative evaluation, it introduced artifacts in the images, particularly noticeable in the front view.

4.4. Applications

GBCN uses a pipeline similar to ECON during reconstruction, allowing quick adaptation to other scenarios. Our method leverages existing image-text texture generation models to recover model textures. Thanks to GBCN's detailed reconstruction, we can directly 3D print the output files without processing. Additionally, we use reconstructed results with texture edits for scene building, as shown in Fig.1 and Fig.3. In summary, GBCN's results can easily extend to real-world applications, offering significant potential in education, gaming, and virtual reality.

5. Conclusion

We propose GBCN, an innovative diffusion model that refines the Brownian Bridge diffusion process and introduces it to human reconstruction. This method delivers notable results, effectively addressing the substantial domain gap between 2D and 3D.The normal map prediction module in GBCN is simple, efficient, and highly portable. Applied to recent works, it demonstrated state-of-the-art performance across multiple metrics and impressive results on natural images.Additionally, we further extend the outputs of GBCN to practical applications, showcasing its strong potential in tasks such as 3D printing and scene construction.

References

[1] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 3

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018.

[4] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 3

[5] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022. 2

[6] Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K Wong. Jiff: Jointly-aligned implicit face function for high quality single view clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2729–2739, 2022. 3

[7] Yukang Cao, Kai Han, and Kwan-Yee K Wong. Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4647–4657, 2023. 3

[8] Jinnan Chen, Chen Li, Jianfeng Zhang, Hanlin Chen, Buzhen Huang, and Gim Hee Lee. Generalizable human gaussians from single-view image. *arXiv preprint arXiv:2406.06050*, 2024. 3

[9] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing controllable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16954–16964, 2023. 2, 3

[10] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. 2, 1

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[13] Ben Fei, Jingyi Xu, Rui Zhang, Qingyuan Zhou, Weidong Yang, and Ying He. 3d gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3

[14] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021. 3, 2

[15] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J Black. Learning disentangled avatars with hybrid 3d representations. *arXiv preprint arXiv:2309.06441*, 2023. 3

[16] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2360–2369, 2021. 5

[17] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8770, 2023. 3

[18] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. 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*, pages 12858–12868, 2023. 3

[19] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 1

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12869–12879, 2023. 3

[21] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33:9276–9287, 2020. 3

[22] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11046–11056, 2021. 3

[23] 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. 2, 3, 7

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 1

[25] 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. 3

[26] 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

[27] Shuo Huang, Zongxin Yang, Liangting Li, Yi Yang, and Jia Jia. Avatarfusion: Zero-shot generation of clothing-decoupled 3d avatars using 2d diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5734–5745, 2023. 3

[28] Shuo Huang, Shikun Sun, Zixuan Wang, Xiaoyu Qin, Yanmin Xiong, Yuan Zhang, Pengfei Wan, Di Zhang, and Jia Jia. Placiddreamer: Advancing harmony in text-to-3d generation. *arXiv preprint arXiv:2407.13976*, 2024. 2

[29] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8974–8985, 2023. 3

[30] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, pages 1531–1542. IEEE, 2024. 2, 3

[31] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 2, 3

[32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[33] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13919–13929, 2021. 5

[34] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 3

[35] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer, 2022. 3

[36] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 3

[37] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 3

[38] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15965–15976, 2023. 2, 3

[39] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 3

[40] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11127–11137, 2021.

[41] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11035–11045, 2021.

[42] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 3

[43] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*, pages 643–653. IEEE, 2019. 3

[44] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023. 2, 3, 1

[45] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 3

[46] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12933–12942, 2023.

[47] Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9110–9121, 2023.

[48] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 3

[49] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. 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*, pages 19711–19722, 2024. 3

[50] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-fidelity clothed avatar reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8662–8672, 2023. 3

[51] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. In *2024 International Conference on 3D Vision (3DV)*, pages 1508–1519. IEEE, 2024. 3

[52] Hongyu Liu, Xuan Wang, Ziyu Wan, Yujun Shen, Yibing Song, Jing Liao, and Qifeng Chen. Headartist: Text-conditioned 3d head generation with self score distillation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[53] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 1

[54] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[55] Yiwei Ma, Zhekai Lin, Jiayi Ji, Yijun Fan, Xiaoshuai Sun, and Rongrong Ji. X-oscar: A progressive framework for high-quality text-guided 3d animatable avatar generation. *arXiv preprint arXiv:2405.00954*, 2024. 3

[56] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3

[57] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[58] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 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*, pages 10975–10985, 2019. 3

[59] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 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*, pages 10975–10985, 2019. 2, 3, 1

[60] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit Bermano, Eric Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. State of the art on diffusion models for visual computing. In *Computer Graphics Forum*, page e15063. Wiley Online Library, 2024. 2

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

[62] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019. 5

[63] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 3

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 1

[65] 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. 2, 3, 7

[66] 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. 2, 3, 7

[67] Akash Sengupta, Thiemo Alldieck, Nikos Kolotouros, Enric Corona, Andrei Zanfir, and Cristian Sminchisescu. Diffhuman: Probabilistic photorealistic 3d reconstruction of humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1439–1449, 2024. 3

[68] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8887–8896, 2023. 3

[69] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023. 2

[70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2, 1

[71] Xiaokun Sun, Zhenyu Zhang, Ying Tai, Qian Wang, Hao Tang, Zili Yi, and Jian Yang. Barbie: Text to barbie-style 3d avatars. *arXiv preprint arXiv:2408.09126*, 2024. 3

[72] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3

[73] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022. 2

[74] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[75] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*, pages 322–332. IEEE, 2020. 3

[76] 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. 2, 3, 7

[77] 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. 2, 3, 6, 7, 1

[78] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 3

[79] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023. 2

[80] Yuanyou Xu, Zongxin Yang, and Yi Yang. Seeavatar: Photorealistic text-to-3d avatar generation with constrained geometry and appearance. *arXiv preprint arXiv:2312.08889*, 2023. 3

[81] Zhiqin John Xu. Understanding training and generalization in deep learning by fourier analysis. *arXiv preprint arXiv:1808.04295*, 2018. 5

[82] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, pages 264–274. Springer, 2019. 5

[83] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-if: Uncertainty-aware human digitization via implicit distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9122–9132, 2023. 2, 3, 7

[84] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6796–6807, 2024. 3

[85] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14718–14727, 2021. 3

[86] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11446–11456, 2021. 3

[87] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023. 3

[88] Jingbo Zhang, Xiaoyu Li, Qi Zhang, Yanpei Cao, Ying Shan, and Jing Liao. Humanref: Single image to 3d human generation via reference-guided diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1844–1854, 2024. 3

[89] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 7

[90] 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. 2, 3, 7

[91] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184, 2021. 2, 3, 7

[92] Yunliang Zhuang, Zhuoran Zheng, and Chen Lyu. Dpfnet: A dual-branch dilated network with phase-aware fourier convolution for low-light image enhancement. *arXiv preprint arXiv:2209.07937*, 2022. 5