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

Supplementary Material

This supplementary material provides details excluded from the main paper due to space limitations. It offers additional information on network model architecture, implementation details, discussions, and more qualitative results as an extension of Sections 3 and 4.

1. Diffusion Models

A T-step Denoising Diffusion Probabilistic Model (DDPM) consists of two processes: the forward process (also referred to as the diffusion process), and the reverse inference process.

The forward process from data $\mathbf{x}_0 \sim q_{\text{data}}(\mathbf{x}_0)$ to the latent variable $\mathbf{x}_T$ can be formulated as a fixed Markov chain:

$$q(\mathbf{x}_1, \ldots, \mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{1}$$

where

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

is a normal distribution, β_t is a small positive constant. The forward process gradually perturbs $\mathbf{x}_0$ to a latent variable with an isotropic Gaussian distribution $p_{\text{latent}}(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The reverse process strives to predict the original data $\mathbf{x}_0$ from the latent variable $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ through another Markov chain:

$$p_\theta(\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}|\mathbf{x}_T) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t). \tag{2}$$

The training objective of DDPM is to optimize the Evidence Lower Bound (ELBO). Finally, the objective can be simplified as to optimize:

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2\right], \tag{3}$$

where $\boldsymbol{\epsilon}$ is the Gaussian noise in $\mathbf{x}_t$, which is equivalent to $\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{x}_{t-1})$, and ϵ_θ is the model trained to estimate $\boldsymbol{\epsilon}$.

Most conditional diffusion models [**?** **?** **?** **?**] maintain the forward process and directly inject the condition into the training objective:

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, y, t)\|_2^2\right]. \tag{4}$$

Since $p(\mathbf{x}_t|y)$ does not obviously appear in the training objective, it is difficult to guarantee the diffusion can finally reach the desired conditional distribution.

Except for the conditioning mechanism, Latent Diffusion Model (LDM) [**?**] takes the diffusion and inference processes in the latent space of VQGAN, which is proven to be more efficient and generalizable than operating on the original image pixels.

2. Implementation Details

2.1. Model Structure

Normal map prediction. We referenced BBDM's [**?**] design and adopted VQGAN as the latent space encoder, with parameters including a downsampling factor of $f = 4$, 8192 tokens in the Vector Quantization (VQ) space, and a latent feature dimension of $d = 3$. All input images were compressed into tensors of size $[\text{batch\_size}, 3, 128, 128]$ and diffused in the latent space.

Additionally, the SMPL-X [**?**] prior extracts multi-scale features through a ResNet encoder and performs cross-attention operations on layers with scales $[1, 4, 8]$ within the U-Net of the diffusion model. We designed a attention decoupling module with a depth of 3. The module employs a multi-head cross-attention mechanism, using the encoded SMPL-X prior image as the query and the intermediate features during the diffusion process as keys and values. We enhance feature extraction and the attention mechanism through positional embedding, incorporating 3D information into the latent space via the decoupling module.

To improve GBCN's performance, we introduced Exponential Moving Average (EMA) and a ReduceLROnPlateau learning rate scheduler during training. The diffusion process across the front and back surfaces is detailed in Tab.1 with different parameter settings.

D

3D Surface Reconstruction. We used d-BiNI from ECON [**?**] as a key step in the reconstruction process. Specifically,we explicitly model the depth-normal relationship using variational normal integration methods.The specific formula is as follows:

$$\text{d-BiNI}(\widehat{\mathcal{N}}_{\text{F}}^{\text{c}}, \widehat{\mathcal{N}}_{\text{B}}^{\text{c}}, \mathcal{Z}_{\text{F}}^{\text{b}}, \mathcal{Z}_{\text{B}}^{\text{b}}) \rightarrow \widehat{\mathcal{Z}}_{\text{F}}^{\text{c}}, \widehat{\mathcal{Z}}_{\text{B}}^{\text{c}}. \tag{5}$$

Here, $\hat{\mathcal{N}}_*^c$ is the front or back clothed normal map predicted by $\mathcal{G}_{\text{F,B}}^{\mathcal{N}}$ from $\{\mathcal{I}, \mathcal{N}^b\}$, and $\mathcal{Z}_*$ is the front or back coarse body depth image rendered from the SMPL-X mesh, $\mathcal{M}^b$.

Max Learning Rate	Min Learning Rate	Factor	Patience	Cool Down	Threshold	eta(front)	eta(back)
1.0e-4	5.0e-7	0.5	3000	2000	1.0e-4	0.3	0.7

Table 1. Parameter configurations for GBCN models.

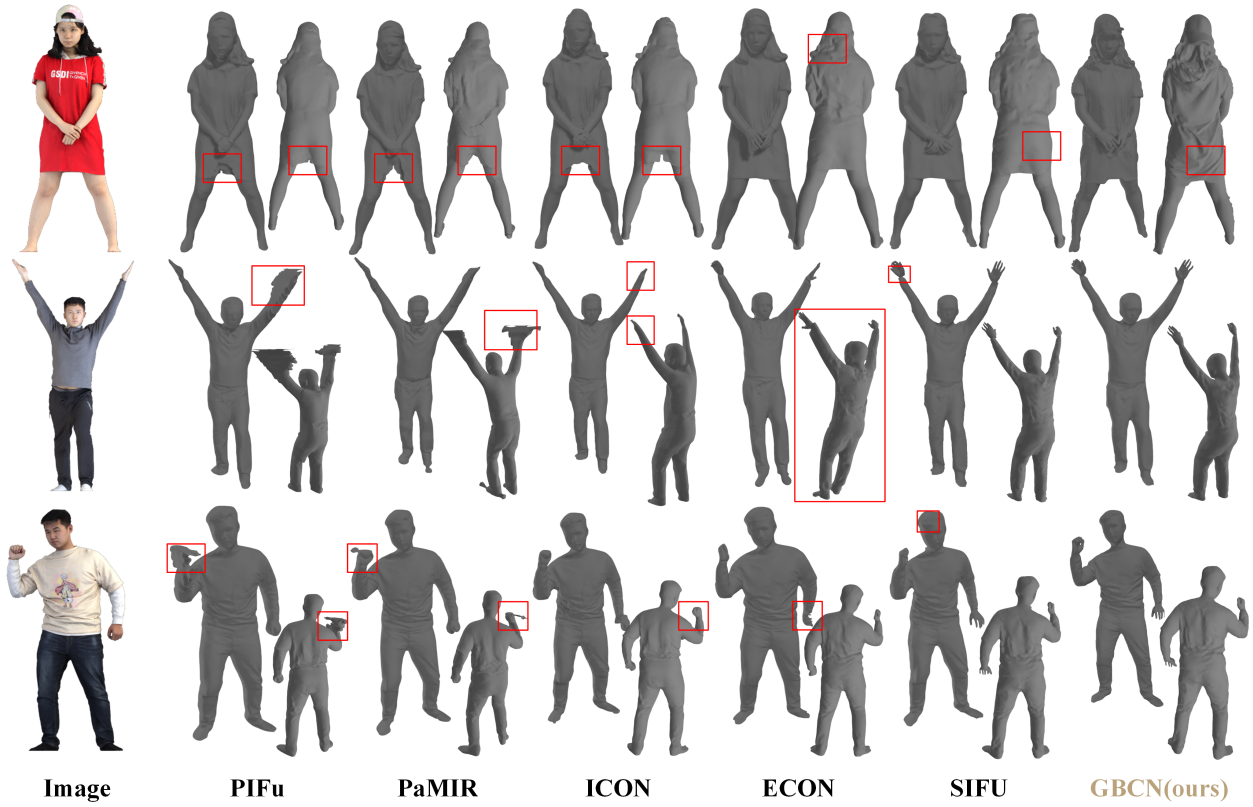

Figure 1. Qualitative comparison of geometry quality on THuman2.0. Please zoom in for details.

our objective function consists of five terms:

$$\min_{\widehat{\mathcal{Z}}_{\mathrm{F}}^{\mathrm{c}},\widehat{\mathcal{Z}}_{\mathrm{B}}^{\mathrm{c}}} \mathcal{L}_{\mathrm{n}}(\widehat{\mathcal{Z}}_{\mathrm{F}}^{\mathrm{c}};\widehat{\mathcal{N}}_{\mathrm{F}}^{\mathrm{c}}) + \mathcal{L}_{\mathrm{n}}(\widehat{\mathcal{Z}}_{\mathrm{B}}^{\mathrm{c}};\widehat{\mathcal{N}}_{\mathrm{B}}^{\mathrm{c}}) + \lambda_{\mathrm{d}}\mathcal{L}_{\mathrm{d}}(\widehat{\mathcal{Z}}_{\mathrm{F}}^{\mathrm{c}};\mathcal{Z}_{\mathrm{F}}^{\mathrm{b}}) + \lambda_{\mathrm{d}}\mathcal{L}_{\mathrm{d}}(\widehat{\mathcal{Z}}_{\mathrm{B}}^{\mathrm{c}};\mathcal{Z}_{\mathrm{B}}^{\mathrm{b}}) + \lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{s}}(\widehat{\mathcal{Z}}_{\mathrm{F}}^{\mathrm{c}},\widehat{\mathcal{Z}}_{\mathrm{B}}^{\mathrm{c}}), \quad (6)$$

where $\mathcal{L}_n$ is the BiNI loss term introduced by BiNI [**?**], $\mathcal{L}_d$ is a depth prior applied to the front and back depth surfaces, and $\mathcal{L}_s$ is a front-back silhouette consistency term.The normal maps we generate do not include occlusions, allowing for straightforward merging of front and back d-BiNI surfaces. To fill in the missing surface information during merging, we use SMPL-X to complete the surfaces, resulting in a complete 3D garment scan.

2.2. Training and Inference

We generate training data using 3D scans from THuman2.0. Each scan is rendered from 36 different angles at a resolution of 512, employing a weak perspective camera that horizontally rotates around the scan, under varying environmental lighting conditions. The model, implemented in PyTorch Lightning, is trained for 10 epochs with a learning rate of 1e-4 and a batch size of 4, over a span of 3 days on a single NVIDIA GeForce RTX 3090 GPU. During training, we set $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 5.0$ for each loss item.

During inference, following SIFU [**?**] and ECON [**?**], we use Rembg for background removal in in-the-wild images and employ PIXIE [**?**] to estimate SMPL-X parameters, refining them further as per [**?**]. We iteratively refine SMPL-X and clothed-body normals for 50 iterations.

3. Limitations and future work

GBCN takes a single image and the estimated SMPL-X model from that image as input, reconstructing the corresponding 3D clothed human mesh. GBCN applies the Brownian Bridge diffusion process to the 3D domain, learn-

Figure 2. Although SiTH performs well on certain images, it still exhibits some instability. Please zoom in for details.

ing 3D details in the latent space and addressing large domain gaps. We output detailed normal maps, resulting in refined reconstruction. However, estimating the SMPL-X body from a single image remains an open problem. Although GBCN applies perturbations to the SMPL-X priors during training to enhance robustness, it still produces poor results for SMPL-X bodies significantly deviating from true values. In our experiments, GBCN used PIXIE to estimate the SMPL-X body. As technology advances, the gap between pose estimation and ground truth is expected to narrow, potentially eliminating this limitation.

Apart from the limitations mentioned, several other directions are crucial in practical applications. GBCN's main contribution focuses on domain mapping from 2D to 3D, neglecting model textures. Existing networks [**? ? ? ? ?**] often use image-text multimodal models to generate prompts for texture generation, but this leads to information loss. Incorporating texture reconstruction into the Brownian Bridge diffusion process would enable mesh texture reconstruction from a single photo.

4. Additional Results

We evaluated the geometric performance of GBCN on the Thuman2.0 dataset and in-the-wild images. As shown in Fig.1, GBCN consistently provides impressive reconstruction results across various scenarios, accurately reproducing complex poses and loose clothing, while capturing detailed geometry in occluded areas. Fig.3 presents qualitative results for in-the-wild images, demonstrating GBCN's robustness in handling complex poses, loose clothing, and substantial geometric detail. Notably, SiTH has achieved notable improvements in metrics. However, the posterior hallucinations produced by its image-conditioned diffusion model are often unstable, requiring manual selection of the generated images. Therefore, its results were not included in comparisons with others. Fig.2 shows the outcomes of its diffusion model.

Figure 3. Qualitative comparison of geometry quality on in-the-wild images. Please zoom in for details.