electronics

MDPI

Article

ATGT3D: Animatable Texture Generation and Tracking for 3D Avatars

Fei Chen † and Jaeho Choi *,†

Department of Electronic Engineering, Jeonbuk National University, Jeonju 54896, Republic of Korea; mophy@jbnu.ac.kr

* Correspondence: wave@jbnu.ac.kr; Tel.: +82-010-5615-2415

† These authors contributed equally to this work.

Abstract: We propose the ATGT3D an Animatable Texture Generation and Tracking for 3D Avatars, featuring the innovative design of the Eye Diffusion Module (EDM) and Pose Tracking Diffusion Module (PTDM), which are dedicated to high-quality eye texture generation and synchronized tracking of dynamic poses and textures, respectively. Compared to traditional GAN and VAE methods, ATGT3D significantly enhances texture consistency and generation quality in animated scenes using the EDM, which produces high-quality full-body textures with detailed eye information using the HUMBI dataset. Additionally, the Pose Tracking and Diffusion Module (PTDM) monitors human motion parameters utilizing the BEAT2 and AMASS mesh-level animatable human model datasets. The EDM, in conjunction with a basic texture seed featuring eyes and the diffusion model, restores high-quality textures, whereas the PTDM, by integrating MoSh++ and SMPL-X body parameters, models hand and body movements from 2D human images, thus providing superior 3D motion capture datasets. This module maintains the synchronization of textures and movements over time to ensure precise animation texture tracking. During training, the ATGT3D model uses the diffusion model as the generative backbone to produce new samples. The EDM improves the texture generation process by enhancing the precision of eye details in texture images. The PTDM involves joint training for pose generation and animation tracking reconstruction. Textures and body movements are generated individually using encoded prompts derived from masked gestures. Furthermore, ATGT3D adaptively integrates texture and animation features using the diffusion model to enhance both fidelity and diversity. Experimental results show that ATGT3D achieves optimal texture generation performance and can flexibly integrate predefined spatiotemporal animation inputs to create comprehensive human animation models. Our experiments yielded unexpectedly positive outcomes.

Keywords: human body; 3D representations; human motion; 3D texture; texture tracking

Citation: Chen, F.; Choi, J. ATGT3D: Animatable Texture Generation and Tracking for 3D Avatars. *Electronics* **2024**, *13*, 4562. https://doi.org/10.3390/electronics13224562

Academic Editor: José Carlos Castillo

Received: 25 October 2024
Revised: 14 November 2024
Accepted: 17 November 2024
Published: 20 November 2024

1. Introduction

Creating 3D virtual humans that replicate real-world scenarios continues to be a long-term objective in computer graphics and vision. The ability to generate and track dynamic 3D textures provides virtual characters with high-quality detail and dynamic consistency, thereby enhancing user immersion. Additionally, in industries such as virtual humans and film animation, the generated high-resolution textures offer innovative solutions for creating virtual characters, making them suitable for applications in advertising, film, and online streaming. Synthesizing texture appearance and tracking models are essential tasks in creating digital humans, both crucial for achieving realistic, video-like authenticity. To produce video-like effects, it is essential to generate complete models to enable the accurate tracking of movements in videos using human textures.

Recent approaches in pose image synthesis tasks [1–7] utilize generative strategies such as GANs [8] or VAEs [9], guided by 2D pose representations or text, and have

demonstrated impressive results. However, these methods directly output 2D images of humans in camera space without generating complete texture maps, limiting their usability in standard 3D animation pipelines that rely on UV [10] texture mapping for texturing 3D meshes. UV mapping refers to the process of projecting a 2D texture map onto a 3D model. Moreover, they often struggle to recover textures from in-the-wild images. Similarly, other works employ neural rendering pipelines [11–15] to generate view-dependent pose avatars but fail to produce 3D texture maps.

Other research focuses on complete texture estimation from a single image [3,16–20] and is capable of recovering 3D texture maps from casual images. Most of these methods employ CNN architectures [16,21] to infer complete 3D texture maps from a single image, sometimes incorporating multiview supervision [20] or Transformer-based architectures [19]. However, these methods are confined to generating textures from images, which limits their applicability in scenarios requiring animated virtual human synthesis. Notably, the limited expressive capacity of the network's latent space often results in low-detail textures. We believe that these limitations obstruct the development of large-scale, high-quality 3D human textures and consider this a significant shortcoming in the field.

Closely related to the task of texture estimation are methods aimed at reconstructing 3D human bodies and appearances from a single image [22–26]. These methods can generate high-fidelity 3D reconstructions, including fine geometric details, but the estimated appearances are often not explicitly baked into consistent UV texture maps. Moreover, they are unable to accurately track textures for animated human models.

To address the challenge of accurately tracking textures for animated human models, we introduce a texture-tracking human animation pose image synthesis method. This methodology predicts nearly complete 3D texture maps and generates images corresponding to human poses [1–7], requiring the use of synthetic 3D human subjects. For instance, Saito et al. [6] encoded partial (i.e., visible) UV space appearances into global latent vectors for the image generator while concurrently correcting the corresponding fields and transferring local surface details to the target pose. They employed a generative technique for the synthesis of full 3D human textures. This approach can determine 3D human textures from single images, which can be directly applied to animated SMPL-X meshes, as shown in Figure 1.

It is used to create facial expressions with textures, dynamics of body parts, and the tracking of hand motions conditioned on the texture maps of the human body and animated model poses. Complete masking of 3D human texture is achieved. Figure 1a displays the texture map and the 3D human model in various poses. Figure 1b presents images of the 3D human body in different poses.

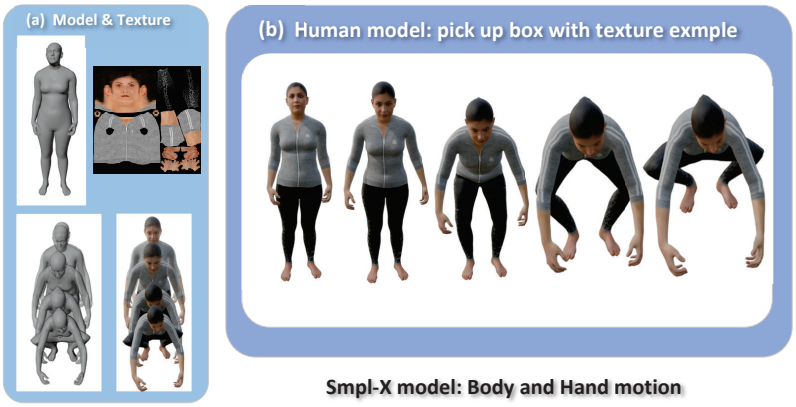

Figure 1. This framework facilitates the generation and tracking of animated textures for 3D virtual images.

We utilize the appearance texture maps from the HUMBI dataset and employ our proposed TED module, which encompasses the generation of appearance texture and the assessment of eye loss in a complete texture map (see Section 4.2), and we also use the 3D human models from the AMASS human motion pose dataset as a baseline for action parameters. Our framework establishes a model for 3D animated texture generation and tracking named ATGT3D. In Section 4.1 depicts the overall structure of ATGT3D. Initially, ATGT3D gathers positional and spatial features of the human eye from an image and generates a complete human texture map using a simple autoencoder through our proposed Eye Diffusion Module (EDM). Detailed information is available in Section 4.1. Subsequently, by converting to capable 3D animation parameters of SMPL-X within the AMASS dataset, the selection of varying forward paths facilitates effective gesture-to-gesture and audio-to-gesture modeling, respectively. Once the reconstructed latent features are obtained, ATGT3D establishes correspondences at different time frames to decode local facial and body poses and decode human texture map parameters from pretrained global motion pose parameters, hence completing the tracking of textures and models.

Overall, our contributions are as follows:

1. We propose the EDM, which enhances 3D human texture parameters through the use of texture seeds and diffusion models, generating nearly complete 3D human texture maps.
2. We introduce a human motion PTDM for mesh-level animatable human model datasets. It is a straightforward and effective texture motion tracking framework that can generate temporally coherent texture motions from a single image.
3. Using the BEAT2 and AMASS datasets, we develop an outstanding human and pose synchronization model using only three seed poses, capable of generating body and facial gestures. This significantly enhances the fidelity and diversity of the results.

2. Related Work

Recently, numerous methods have achieved notable success in image pose synthesis tasks [1–7]. Other studies focus on complete texture estimation from a single image [3,16–20], demonstrating effectiveness in reconstructing 3D texture maps from images. In this discussion, we explore the most relevant human texture generation methods for 3D human models, primarily focusing on texture tracking.

2.1. Model Transformation

The parameterized 3D body model SMPL [27] represents the human body. It can be conceptualized as a base model with a series of deformations applied. Based on these deformations, Principal Component Analysis (PCA) [28] is performed to derive low-dimensional shape parameters that characterize the body's shape. Meanwhile, the body's pose is described through a kinematic tree, which outlines the rotational relationship between each joint and its parent node in the tree. This relationship is represented as a three-dimensional vector, and the local rotational vectors of each joint collectively define the pose parameters of the SMPL model.

Although the SMPL model is widely used, it lacks specific features such as articulated hands and expressive faces. The SMPL-X model [29] addresses these deficiencies by integrating hands and faces; however, while SMPL-X extends SMPL technology, the two models are not fully interchangeable. Notably, despite the shape and pose parameters of SMPL and SMPL-X looking very similar, they are not directly transferable. In particular, joint positions in SMPL-X differ from those in SMPL, making the pose (θ) parameters incompatible.

In this section, we outline a tool for converting parameters between these models. This process involves fitting one model to the other to recover the corresponding parameters. Specifically, it involves establishing a mapping between the SMPL and SMPL-X models by locating the nearest point on the SMPL mesh for each SMPL-X vertex. The process includes storing index_i^t, the position of the nearest point within a triangle, and $[a_i, b_i, c_i]$, the barycentric coordinates of the nearest point relative to the SMPL triangle.

SMPL and SMPL-X share identical topology up to the neck, and the coordinate centers of these points are represented as [1.0, 0.0, 0.0]. We also maintain a mask of valid vertices to exclude mismatched points between the two meshes, such as the eyeballs or inside of the mouth. Having established these correspondences between the models, we can align the SMPL-X annotations with the SMPL ones, constructing a proposed SMPL annotation based on the SMPL-X topology, as shown in Equation (1):

$$v_i^{SMPL\text{-}X} = a_i * v_{f_i^0}^{SMPL} + b_i * v_{f_i^1}^{SMPL} + c_i * v_{f_i^2}^{SMPL}, \tag{1}$$

With a mesh conforming to the SMPL-X topology, our goal is to determine the SMPL-X parameters that best describe this mesh: the pose parameters θ, shape parameters β, expression parameters ψ, and translation parameters γ. We use an iterative optimization scheme to recover these parameters for the model.

SMPL-X is an animatable parametric 3D body model that integrates the body, face, and hands. It consists of N = 10,475 vertices and $K = 54$ joints. Utilizing the shape parameters β, pose parameters θ (which encompass body joint poses θ_b, jaw pose θ_f, and hand pose θ_h), and expression parameters ψ, the SMPL-X model depicts the human body as $M(\beta, \theta, \psi)$, as illustrated in Equation (2):

$$\begin{aligned} \mathbf{M}(\beta, \theta, \psi) &= \mathcal{W}(\mathbf{T}(\beta, \theta, \psi), J(\beta), \theta, W) \\ \mathbf{T}(\beta, \theta, \psi) &= T + B_s(\beta) + B_e(\psi) + B_p(\theta), \end{aligned} \tag{2}$$

where T represents a mean shape template, and B_s, B_e, and B_p correspond to shape, expression, and pose blend shapes, respectively. W is the linear blend-skinning function that adapts $\mathbf{T}(\beta, \theta, \psi)$ to the target pose θ, utilizing the skeleton joints $J(\beta)$ and skinning weights $W \in \mathbb{R}^{N \times K}$.

2.2. Texture Repair

Latent diffusion models (LDMs) [30] decompose the image formation process into a continuous sequence of denoising autoencoders. This paper explores the in-painting capabilities of latent diffusion models for images. However, adapting diffusion models to three-dimensional human bodies introduces two significant challenges: spatial regularization for multiview consistency and 3D perception for converting 2D images into 3D tasks.

Initially, we train a domain-specific diffusion model capable of generating unwrapped 3D human body textures, simultaneously learning multiview consistency. To enable the 3D perception required for 2D-to-3D tasks (e.g., deriving 3D textures from monocular input), we identify pixel-to-surface correspondences [31], projecting image pixels onto an incomplete 3D texture map. By capitalizing on the inherent 2D structure within the domain-specific diffusion model, we effectively correct the incomplete 3D texture map.

To synthesize a complete 3D texture from images in the HUMBI dataset, we initially train a specialized eye-domain diffusion module (EDM) to generate clear, unwrapped 3D textures of human eyes. Subsequently, to provide the necessary 3D perception for 2D-to-3D tasks (e.g., creating 3D textures of the human body), we locate the 3D texture map within the HUMBI dataset. We compute the pixel-to-surface correspondences [31], specifically those related to the eyes, to project image pixels and compile a complete 3D texture map. By exploiting the inherent 2D structure of the eye-specific diffusion model, we amend the incomplete 3D texture map.

Our approach is also related to body synthesis methods driven by text descriptions [32–34]. These methods utilize GANs [8], VAEs, or diffusion models [1] combined with comprehensive visual language pretraining models (like CLIP [35]) to regulate the output. Although these models can create 3D textured human bodies and animate them based on text input, they face challenges in generating consistent texture maps or adeptly adapting to local views for texture estimation in images. We propose a pipeline built on the novel Eye

Diffusion Module, capable of recovering high-quality textures and performing natural human model fitting operations.

2.3. Motion Tracking

To initialize the SMPL-X body shape and pose parameters, we employ MoSh++ [36,37] using data derived from motion capture markers. Given the captured marker positions $\mathbf{S} \in \mathbb{R}^{T\times K\times 3}$, predefined marker position offsets $\mathbf{d} \in \mathbb{R}^{K\times 3}$, and a user-defined vertex-to-marker function $(\mathcal{H})$, our objective is to accurately determine body shape $\beta \in \mathbb{R}^{300}$, pose $\theta \in \mathbb{R}^{T\times 55\times 3}$, and translation parameters $\gamma \in \mathbb{R}^{T\times 3}$. The optimization involves a differentiable surface vertex mapping function $\mathcal{S}(\beta,\theta,\gamma)$ and a vertex normal function $\mathcal{N}(\beta,\theta_0,\gamma_0)$. For each frame, the latent markers $\tilde{\mathbf{m}} \in \mathbb{R}^{T\times K\times 3}$ are computed as $\tilde{\mathbf{m}}_i \equiv \mathcal{S}_{\mathcal{H}}(\beta,\theta_i,\gamma_i) + \mathbf{d}\mathcal{N}_{\mathcal{H}}(\beta,\theta_i,\gamma_i)$. An evaluation formula is proposed, where the validation of the action is performed using the average minimum difference across three frames in the model action sequence, as explained in Section 4.2.

$$L_p = \frac{1}{frame\ n}\sum_{i=i}^{n} S_H(\beta,\theta_i,\gamma_i) + dN_H(\beta,\theta_i,\gamma_i), \tag{3}$$

2.4. Literature Review

The application of diffusion models in 3D texture generation and tracking has garnered increasing attention. By incrementally adding detail noise during the generation of complex 3D textures, diffusion models can effectively enhance texture precision and realism.

GAN [8] and VAE [32] methods exhibit certain limitations in dynamic scenes. While GANs can generate realistic 2D textures, they struggle to maintain texture consistency in 3D animated scenes, particularly during rapid movements and multiangle transitions, often resulting in texture blurring and distortion. Although VAEs demonstrate stability in producing diverse textures, they fall short in detail resolution compared to diffusion models and struggle to maintain texture continuity amid complex pose variations. These limitations restrict the applicability of GANs and VAEs in dynamic 3D scenarios. To address these challenges, ATGT3D introduces the EDM and PTDM to achieve higher-quality dynamic texture generation. The EDM focuses on generating high-resolution eye texture details, leveraging the progressive refinement capabilities of diffusion models to ensure realism across multiangle and dynamic scenes. The PTDM employs temporal modeling to synchronize pose and texture tracking, effectively overcoming the texture consistency issues encountered by traditional methods in fast-motion and complex animated scenes. In comparison, ATGT3D demonstrates superior robustness and detail accuracy in dynamic scenarios, offering an innovative solution for virtual human and 3D animation generation.

3. Datasets and Preprocessing

In this section, we introduce the HUMBI [38] dataset used for acquiring texture images, the AMASS [37] dataset leveraged for integrating texture with human models, and the BEAT2 [32] dataset employed for evaluating the effects.

3.1. HUMBI Dataset

The HUMBI dataset [38] is an extensive multiview dataset designed to capture human body expressions in natural clothing. It was captured at a frequency of 60 Hz with 107 GoPro HD cameras from 772 different subjects. It is organized by subject, with each subject having four expression sessions. Each session features a sequence of frames (temporal instances), with each frame containing up to four representations: multiview images, 3D keypoints, 3D meshes, and appearance (texture) maps. We utilize the frontal appearance maps from HUMBI for texture mapping in this study, and we take the 32nd appearance of HUMBI with multiple angles as the frontal image.

3.2. AMASS Dataset

AMASS [37] is a comprehensive and diverse database of human motion that combines 15 different optical marker-based motion capture datasets into a single framework and parameterization. We utilized a novel dataset of 4D body scans concurrently recorded with marker-based motion capture to evaluate MoSh++ [37] and fine-tune its hyperparameters. The unified representation delivered by AMASS simplifies its application in animation, visualization, and the generation of deep learning training data. This dataset is significantly richer than previous collections of human motion, collecting over 40 h of motion data from more than 300 subjects and over 11,000 actions.

AMASS consists of 40 h of motion capture data, 344 subjects, and 11,265 actions. The original datasets constituting AMASS feature between 37 and 91 variably distributed motion capture markers. Each frame in AMASS includes SMPL 3D shape parameters (16 dimensions), DMPL soft tissue coefficients (8 dimensions), and complete SMPL pose parameters (159 dimensions), which encompass hand joints and global body translation. We utilize the SMPL to SMPL-X model to derive the animation features of the SMPL-X model within the AMASS dataset.

To streamline data loading and balance the proportions between the model and comparison datasets, we selected subsets from the five datasets in AMASS that feature the highest number of motion instances and the shortest recording time. The selection strategy is illustrated in Figure 2, and Table 1 detailed the motions and minutes for the top five subsets of the dataset.

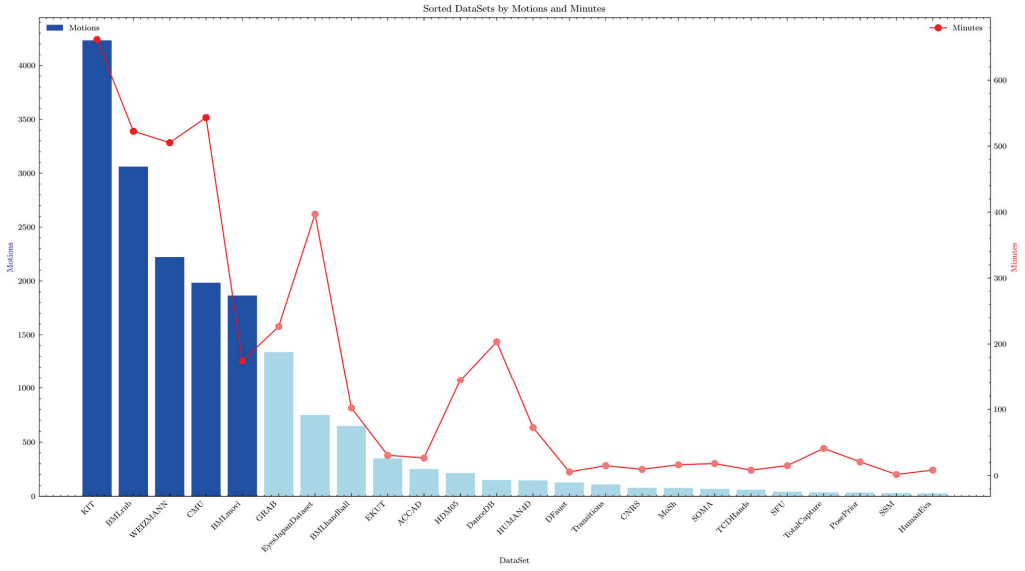

Figure 2. The AMASS dataset is sorted based on the attributes with the most actions (motions) and the least time (minutes). The light blue bars represent subsets of the dataset not utilized in the study, dark blue bars remaining subsets were selected for evaluation and experimentation.

Table 1. Details of the top five subsets, categorized by number of poses and minutes, for selected subsets of the AMASS dataset.

Dataset	Subjects	Motions	Minutes
KIT [39]	55	4232	661.84
BMLrub [40]	111	3061	522.69
WEIZMANN [39]	5	2222	505.35
CMU [41]	96	1983	543.49
BMLmovi [42]	89	1864	174.39

Only the first 5 sub-datasets are selected here.

"The KIT [39,43] Bimanual Manipulation Dataset" and "WEIZMANN" [44] form a multimodal dataset that emphasizes learning tasks involving bimanual manipulation. The dataset includes detailed whole-body motion data, complete configurations of both hands, and 6D poses and trajectories of all associated objects. This dataset covers 12 everyday bimanual household activities performed by two healthy subjects. Each activity features significant intra-action variations, with three repetitions of each variation, resulting in a total of 588 recorded demonstrations. A total of 21 household items were utilized to perform various actions. Furthermore, researchers have developed tools and methods for standardizing the representation and organization of multimodal sensor data in large-scale human motion databases. They have also extended the Master Motor Map (MMM) framework to allow the mapping of collected demonstrations to a human reference model, segmentation, and annotation of recorded manipulation tasks. These tools are publicly accessible in the KIT Whole-Body Human Motion Database.

BMLrub [40] transforms biological motion, such as human gait, into a format analyzable by linear statistical methods and pattern recognition techniques. Using gender classification as an example, the study constructs a simple classifier and benchmarks it against psychological data from human observers. The findings suggest that the dynamic aspects of motion carry more information about gender than do structural cues influenced by motion. This dataset has significantly advanced gait recognition research by introducing a novel method for analyzing and understanding human gait patterns.

The MoVi dataset, developed by Saeed Ghorbani and colleagues and published in *PLoS One*, is an extensive, multipurpose dataset encompassing human movement and video. Noteworthy for its multimodality, the MoVi dataset incorporates optical motion capture, video data, and IMU data. It furnishes researchers with extensive resources for exploring human pose estimation, action recognition, motion modeling, gait analysis, and body shape reconstruction.

3.3. BEAT2 Dataset

To generate full-body human poses from audio and masked gestures, EMAGE [32] introduced the BEATX dataset, which combines MoShed SMPL-X body parameters and FLAME head parameters to enhance the modeling of head, neck, and finger movements. As a high-quality, community-standardized 3D motion capture dataset, BEATX stands out for its detail. The subset BEAT2, which is part of this dataset, comprises 60 h of data that merge SMPL-X limb parameters with FLAME facial parameters, specifically targeting the modeling of head, neck, and fingers movements. In this methodology, the full-body BEAT2 dataset is employed as validation data for human texture motion tracking.

4. Generation Architecture

Our method comprises two components: texture generation (EDM) and the tracking and matching of texture with modeled action poses (PTDM). High-quality textures for matching the human animation model can be found in the texture maps in the appearance section of the HUMBI dataset; texture maps are compared, revealing that improvements are necessary to satisfy the requirements for realistic human reconstruction. Notable issues include ghosting in the reconstructed eyes and pose tracking inaccuracies.

During the training process of the ATGT3D model, the learning rate was set to 2.5×10^{-4}, demonstrating a well-balanced performance and enabling the model to converge within a reasonable number of epochs. We employed the Adam optimizer for parameter updates, which provides robust handling of high-dimensional parameters. The batch size was set to 64, fully leveraging hardware resources. We recommend using a GPU with at least 24 GB of VRAM to support the batch size requirement of 64. Additionally, distributed training can further enhance training speed. For different datasets or application scenarios, adjustments to the learning rate and batch size may be necessary.

4.1. Eye Diffusion Module (EDM)

To address the ghosting issue in the reconstructed eyes, a common method [19,21] begins by constructing an incomplete texture map using a coarse geometric proxy [27] and inferring pixel-to-surface correspondences [45]. This is followed by employing an image-to-image translation framework to repair or estimate the incomplete textures. Nonetheless, the limited detail expressiveness of existing methods prompted us to propose a pipeline based on the SMPL-X model and the image diffusion model [30].

Our proposed pipeline utilizes the image diffusion model [30] to recover high-quality textures, as visualized in Figure 3. Subsequent sections detail the selection of the texture map and its application in estimating complete 3D textures from SMPL-X using monocular RGB images.

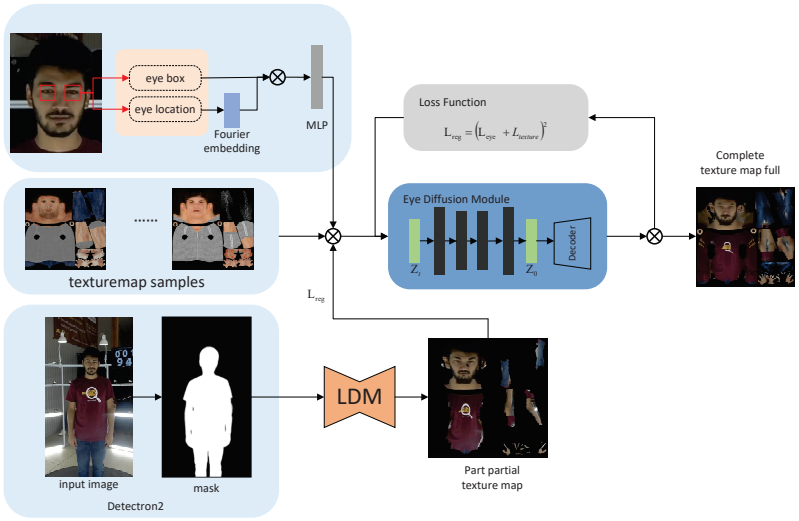

Figure 3. Overview of the proposed method for texture recovery estimation from a single image.

We explored how LDM [30] adapts to the 3D human appearance in the texture generation section and how to produce comprehensive 3D texture maps, including eyes, through the 3D-to-2D parameterization of surfaces (i.e., UV mapping of the mesh surfaces). By directly managing the UV maps, we encode the 3D appearance of the human body onto 2D images, thus enhancing the potential of LDM models for appearance synthesis, restoration, and manipulation.

In generating complete texture maps, we implement category-specific prior retention loss, enabling the model to synthesize target subjects with minimal training samples. We fine-tune the model using one UV texture map from SMPL-X for 1500 iterations.

The process of eye texture reconstruction in the 3D human model is depicted in Figure 4, where initially, Detectron2 [46] extracts the human body region from the input image, creating the corresponding segmentation mask. Subsequently, the Latent Diffusion Model (LDM) processes this image to generate a partial human texture map. This partial texture map lacks the detailed features in the eye area and some edges that are present in a complete texture map, which we aim to produce with a focus particularly on the eye area.

Initially, the eye area is detected and located, yielding the eye frame and position. This position is then subjected to Fourier Embedding to capture its spatial information, which is fed into a multilayer perceptron (MLP) [47] to extract features. These features are combined with the generated texture map and input into a diffusion model targeting the eye area specifically. This model enhances the eye area details by learning from sample texture map data.

The EDM not only reconstructs the eye area details but also ensures the uniformity of the overall texture. Lastly, the outputs from the eye diffusion model are merged with the partial texture map to produce a complete human texture map. This achieves the goal of

detail recovery in the eye area of the 3D human model, with the diffusion model playing a crucial role in enhancing the eye details while maintaining overall texture consistency.

This loss function is specifically engineered to reconstruct fine-grained details in the eye region, focusing on capturing subtle features such as eye contours, iris details, and variations in lighting that are essential for achieving photorealism. The formula for the loss function is presented in Equation (4):

$$L_{\text{eye}} = \mathbb{E}_{(x,y)\in R_{\text{eye}}}\left[\| I_{\text{real}}(x,y) - I_{\text{pred}}(x,y)\|_2^2\right], \tag{4}$$

where $(x,y) \in R_{\text{eye}}$ are the pixel coordinates within the eye region, where R_{eye}, $I_{\text{real}}(x,y)$ denotes the intensity of the pixel in the actual eye region and $I_{\text{pred}}(x,y)$ represents the predicted pixel intensity of the generated eye texture. Moreover, notation $\|\cdot\|_2^2$ indicates the squared error norm. This function measures the pixel-wise difference between the real and the generated eye textures, ensuring faithful reconstruction of intricate details in the eye area.

The Overall Texture Consistency Loss Function, denoted as L_{texture}, ensures that the texture remains consistent throughout the entire 3D model, thus preventing discontinuities or artifacts at the UV map boundaries. It promotes smooth transitions between different texture map regions.

$$L_{\text{texture}} = \mathbb{E}_{(x,y)\in T}\left[\| I_{\text{real}}(x,y) - I_{\text{pred}}(x,y)\|_2^2\right] + \lambda \cdot \mathcal{L}_{\text{smooth}}, \tag{5}$$

Here, $(x,y) \in T$ represent the pixel coordinates within the entire texture map T, $I_{\text{real}}(x,y)$ is the ground-truth pixel intensity in the texture map, and $I_{\text{pred}}(x,y)$ is the predicted pixel intensity. The term L_{smooth} is a regularization component that ensures smooth transitions across adjacent UV regions, while λ is a weighting factor that controls the influence of the smoothness term. This function evaluates the similarity between predicted and real textures across the entire map, maintaining smoothness at UV seams or borders.

To combine both the eye detail and texture consistency losses into a unified framework, we define the total regularization loss L_{reg} as follows:

$$L_{\text{reg}} = (L_{\text{eye}} + L_{\text{texture}})^2, \tag{6}$$

In Equation (6), the square of the sum of the L_{eye} and L_{texture} terms, represented as $L_{\text{reg}} = (L_{\text{eye}} + L_{\text{texture}})^2$, is used to penalize larger deviations more heavily. By squaring the sum, the function increases the loss value more significantly for larger errors, which encourages the model to reduce both L_{eye} and L_{texture} as much as possible. This approach is often used in loss functions to emphasize greater accuracy, as even slight improvements in reducing the overall error become more impactful when squared. Additionally, squaring the total regularization loss enforces a more unified and cohesive structure by strongly discouraging inconsistencies in both eye details and texture consistency. This combined loss enhances both precise eye texture reconstruction and consistent overall texture mapping, thereby ensuring high visual fidelity for the 3D human model's appearance.

4.2. Pose Tracking Diffusion Module (PTDM)

The SMPL-X model, along with the pose parameters, is updated for each frame. Thus, we represent these parameters for the current frame f as $M_f(\beta, \theta_f)$. The pose parameters and the animated avatar mesh for the current frame are represented by θ_f and M_f, respectively.

The SMPL-X model is described in Equation (1), where W denotes the standard linear blend skinning function. This function calculates the output 3D mesh based on the pose joint i, pose parameter a, and blend weight v. The skinned template T is delineated in Equation (2), where T represents the template's average shape, and B_S, B_E, and B_P symbolize the blend shape functions for shape, facial expression, and pose, respectively.

B_S is computed using the shape parameter β and PCA-based [28] shape weights S; B_E is derived from the facial expression parameter ψ and the PCA-based expression weights e; and B_P is derived from the pose parameter θ using the PCA-based pose weights P. The PCA basis is extracted from the samples via PCA.

Rendered normal images can be serve as shape encodings for diffusion models to assist geometric synthesis [48]. Despite this, the method may encounter challenges in achieving perfect consistency between geometry and texture. To mitigate this, we computed the interpolation loss between the latent values of normal images and color images at various times. We refer to this module as PTCN, which is presented in Equation (5):

$$\nabla_\gamma \mathcal{L}_c(\phi, x) = \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}_\phi(\tilde{z}_t; y, t) - \epsilon)\frac{\partial \mathcal{N}}{\partial \gamma}\frac{\partial z}{\partial \mathcal{N}}\right], \tag{7}$$

where $\gamma = \{\beta, \psi, \mathrm{D}\}$ are the geometry-related parameters and $\tilde{z} = \alpha z^{\mathcal{I}} + (1-\alpha)z^{\mathcal{N}}$ denotes the interpolated latent code, with $z^{\mathcal{I}}$ and $z^{\mathcal{N}}$ representing the latent codes for the RGB and normal images, respectively.

The latent features of the current frame are modulated by learnable quantization scalers, as depicted in Figure 4.

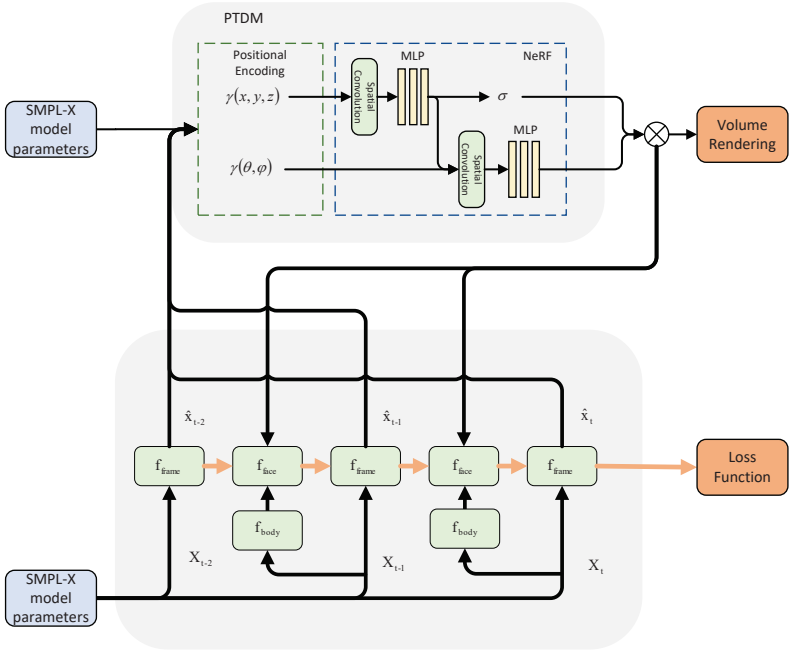

Figure 4. Complete texture tracking in our method to match 3D human models.

As shown in Figure 4, the input pose parameters are first normalized to serve as inputs for generating the initial 3D model. Sequential modeling is then applied to capture dynamic relationships across multiple frames, ensuring motion continuity. The generated pose sequence is mapped onto the texture map. Through the motion–texture synchronization module, the system ensures precise alignment of motion and texture, even during rapid movements or complex poses, ultimately producing high-quality 3D animation textures with temporal consistency.

The surface vertex mapping function and the vertex normal function are essential control points for texture mapping. The pose parameters and translation parameters are crucial in motion training along the temporal relationship axis.

Given the captured marker positions $\mathbf{m} \in \mathbb{R}^{T\times K\times 3}$, predefined marker position offsets $\mathbf{d} \in \mathbb{R}^{K\times 3}$, and user-defined vertex-to-marker function H, our objective is to determine the body shape $\beta \in \mathbb{R}^{300}$, pose $\theta \in \mathbb{R}^{T\times 55\times 3}$, and translation parameters $\gamma \in \mathbb{R}^{T\times 3}$. The optimization utilizes a differentiable surface vertex mapping function $\mathcal{S}(\beta, \theta, \gamma)$ and a

vertex normal function $\mathcal{N}(\beta, \theta_0, \gamma_0)$. For each frame, the potential markers $\tilde{\mathbf{m}} \in \mathbb{R}^{T \times K \times 3}$ are calculated as follows:

$$\tilde{\mathbf{m}}_i \equiv \mathcal{S}_{\mathcal{H}}(\beta, \theta_i, \gamma_i) + \mathrm{d}\mathcal{N}_{\mathcal{H}}(\beta, \theta_i, \gamma_i), \tag{8}$$

We concentrate on optimizing and fixing $\mathbf{d}$ and β for three frames, then optimizing θ_i and γ_i for $i \in (1:T)$ by minimizing the loss term $|\tilde{\mathbf{m}}_i - \mathbf{m}_i|^2$, which encompasses the data term, surface distance energy, and markers. This process begins with regularization, pose and shape priors, velocity constancy, and soft tissue terms. The overall objective function balances accuracy and plausibility through a weighted sum of these terms.

Subsequently, we employ the Gaussian truncation method, adjusting all data points that fall outside the 3σ range to conform to the 3σ threshold and integrate with the adjacent three frames. However, due to the head markers being worn on a helmet, MoSh++ sometimes produces unnatural head shapes and finger poses.

This diagram illustrates a process for generating dynamic textures on 3D human models, integrating a parametric human model (SMPL-X), Neural Radiance Fields (NeRFs), and time-series modeling. This comprehensive process facilitates texture mapping of 3D human models in dynamic scenes via multistage feature fusion and rendering.

Initially, the input and parameterization stage employs the SMPL-X model to parameterize the facial and body features of the human body, denoted by f_{face} and f_{body}, respectively. These features are processed by an encoder module that captures the geometric and pose information of both the face and body. These encoded features, along with the 3D coordinates and viewpoint data of the scene, are subsequently fed into the Neural Radiance Fields module (PTNeRF).

Within the PTNeRF module, the input features are transposed into a high-dimensional space using positional encoding, denoted as $\gamma(x, y, z)$ and $\gamma(\theta, \phi)$. These positional encodings characterize both the points in 3D space and their viewing directions. Following this, the features undergo processing through spatial convolutions and multilayer perceptrons (MLPs) [47], predicting the voxel density σ and the color information for each point. The NeRF output then creates the density and color fields of the 3D human body through volume rendering, providing an exhaustive representation of the character.

The time series modeling component utilizes sequential frame data over time to depict dynamic behavior. The input features for the time series include data from both the current and previous frames X^{t-2}, X^{t-1}, which serve to predict the current frame X^t. Each frame incorporates three distinct features: the global frame feature f_{frame}, facial feature f_{face}, and body feature f_{body}.

These features are cyclically linked along the time axis and undergo updates with each frame. This method facilitates smooth transitions and ensures consistent texture modeling over time.

In the final texture fusion and rendering step, the previously generated EDM results are employed as the initial texture, merging features from the current time frame to maintain both spatial and temporal texture consistency. By feeding these features into the NeRF module, the final 3D human model is created through volume rendering. This output model boasts intricately detailed texture mapping in dynamic scenarios, achieving superior quality in surface texture restoration on the human body. Time series modeling: Dynamic texture mapping is realized by sequentially updating features, capturing the dynamic traits of the character. Neural Radiance Fields (NeRFs): NeRFs serve to create a comprehensive representation of the 3D human model, while volume rendering techniques are employed to enhance detail and depth. Texture consistency: By integrating the earlier EDM results, we ensure consistency and high fidelity across different time frames for the textures produced. This process achieves realistic rendering of 3D human textures in dynamic settings by extensively leveraging parametric human models, NeRFs, and time series data.

4.3. ATGT3D Network Architecture

The temporal sequence modeling in the EDM and PTDM draws upon classical time series techniques, including the VAE method, which performs well in modeling continuity for dynamic pose and texture generation. VAE is also widely used in the generation and prediction of sequential data, and its stability in producing continuous frame sequences provides a solid theoretical foundation for our approach.

In Figure 5, our method utilizes SMPL-X [29] to generate 3D human models, specifically focusing on the face and body parts of SMPL-X. We employ model action prompts to track body movements consistent with real-world scenarios. A key component in this process is the shape encoding within the rendered image diffusion model, crucial for integrating body prompt features. The Pose Tracking Computation Network (PTCN) calculates potential interpolation loss values for different images across various times. The calculation method for PTCN is detailed in Equation (5) of Section 4.2. The Full Texture Module is tasked with restoring high-quality textures using the image diffusion model. We apply benchmark textures from the texture map to restore texture images in the HUMBI dataset using the image diffusion model. These texture images include those of the human eye.

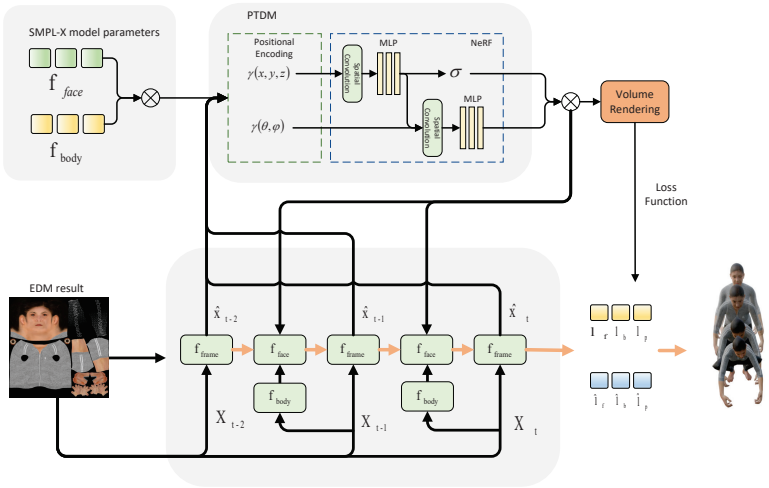

Figure 5. Texture generation as well as tracking and matching graphs of textures with modeled action poses.

5. Experiments and Results

We evaluated the model's capability to extract image textures for each subject in the HUMBI [38] dataset. Animation features from the SMPL-X [29] model were derived from the AMASS [37] dataset, with the BEAT2 [32] dataset serving for comparison. Expressive full-body animated avatars were created, and their quality, as well as texture–geometry consistency, was assessed. Results were documented with a training, validation, and test split of 85%, 7.5%, and 7.5%, respectively. Finally, the effectiveness of each module was analyzed.

To comprehensively evaluate the performance of ATGT3D, we selected multiple baseline methods, such as GAN and VAE-based 3D texture generation approaches, for comparison. Table 2 presents the quantitative results of different methods regarding texture quality and consistency. In Figure 6, we showcase high-resolution images of the texture details generated by each method, with annotated key areas to assist readers in understanding the distinctions between approaches. Additionally, we further assessed each method's performance in complex poses by measuring pose and texture consistency errors. Figure 7 illustrates the error variations across dynamic scenes, showing that ATGT3D demonstrates superior robustness in long-term dynamic tracking compared to baseline methods.

In our experiments, we utilized the Ubuntu 20.04 operating system, the PyTorch 2.1.0 framework, and a server equipped with two Nvidia RTX 4090 GPUs and 64 GB of memory. After 15,000 training iterations over 50 h, we obtained the final results.

Figure 6. Part (**a**) depicts the texture map processed using the image diffusion model to recover high-quality texture. Part (**b**) shows the untrained texture map. Clearly, the clarity of the eye part is superior in image (**a**) compared to image (**b**).

Figure 7. Examples ofmultiple actions across multiple datasets. From top to bottom: natural human postures of various actions for (**a**) AMASS jump Model and AMASS jump Texture, (**b**) AMASS pick-up Model and AMASS pick-up Texture, and (**c**) EMAGE Model and EMAGE Texture.

5.1. Eye Reconstruction

Enhancing eye details is crucial in this processing procedure as eyes are the most expressive part of the face. Advanced texture processing techniques significantly improved the realism of the 3D model, particularly when viewed up close, making it essential for enhancing the expressiveness of virtual characters in real-world applications.

We address the issue of eye ghosting after reconstruction by introducing the LDM [30] model for appearance synthesis and restoration in Section 4.1. This model employs category-specific prior preservation loss during the generation of complete texture maps, enabling synthesis of the target subject with a minimal number of training samples. The results are demonstrated in Figure 6.

Figure 6 illustrates the significant differences in the 3D human model before and after facial texture processing, emphasizing the eye details. Figure 6a demonstrates the 3D model without EDM processing, where the facial area, especially around the eyes, lacks clarity, resulting in a blurred overall appearance that reduces the model's realism, particularly in high-precision areas such as the face.

Conversely, Figure 6b displays the model after EDM processing, where eye details are significantly enhanced, exhibiting more defined textures, realistic contours, improved

lighting, and material representation. This enhancement elevates the visual effects of the face and renders the overall model more vivid and realistic.

5.2. Motion Texture Reconstruction

We utilized the AMASS [37] and BEAT2 [32] model evaluation reports. The AMASS model undergoes rigorous screening and quality control, providing a rich and diverse sample of human motion ideal for evaluating human posture and shape estimation methods. BEAT2 features body data refined by animators and hand data filtered by annotators. The texture tracking effects of the 3D human model under various actions are illustrated in Figure 7, which includes the jumping motion from the AMASS dataset (a), picking action from the AMASS dataset (b), and speaking action from the EMAGE method (c). To ensure consistency, we converted the models in the AMASS dataset from SMPL to SMPL-X, optimized their textures using the Eye Diffusion Module (EDM), and trained them through the Pose Tracking and Diffusion Module (PTDM).

We compared different action sequences of SMPL-X 3D human models from the AMASS dataset to those from the BEAT2 dataset. Each model was evaluated on movements of fixed duration to determine the sequence that best captured quality. The results are displayed in Figure 7.

In part (a), the model shows precise posture tracking during jumping motions. Even with significant movement, the texture details remain clear, effectively conveying natural folds in clothing and muscle lines. The model exhibits strong dynamic consistency, with overall movements appearing coherent and smooth. In part (b), the picking-up motion demonstrates the model's performance during relatively subtle motion changes. The waist bending and arms extending are accurately tracked. The texture maintains high fidelity during dynamic transitions, and the alignment of facial and clothing textures corresponds well with the movements. Part (c) features EMAGE action with relatively smaller movements; however, the model's tracking remains accurate, particularly in expressing hand movements naturally. Although movements in this section are more static compared to parts (a,b), texture mapping maintains clear details, such as facial expressions and body contours.

Overall, the comparison of the three sections demonstrates the PTDM's exceptional performance across various motion types, maintaining consistent model posture during significant movements and exhibiting a high texture fidelity in subtle motions. Clearly our model delivers robust texture generation and model tracking across different datasets.

FGD [49] $(X, \hat{X})$ is defined as the distance between the Gaussian means and covariances of the latent features of human gestures X and the produced gestures $\hat{X}$. FGD is used to assess the realism of body postures. The Frechet Inception Distance (FID) [50] is used to evaluate the distributional discrepancy between generated and real movements, where a lower FID score indicates better performance. The FID metric employs a geometric feature that produces a Boolean vector representing geometric relationships between specific body keypoints in the motion sequence. Furthermore, a kinematic feature extractor converts velocity and acceleration data into the motion sequence. We calculate Diversity [51] by determining the average L1 distance between multiple body gesture clips. To assess Diversity, as outlined in [51], we calculate the average L1 distance between multiple body gesture clips. For hand data, we compute the vertex Mean Squared Error (MSE) [52] to quantify positional discrepancies and the vertex L1 error, also known as LVD [53], between the ground truth (GT) and the generated hand vertices. Additionally, we calculate landmark velocity differences to evaluate the velocity discrepancies between predicted ground truth (p-GT) and generated hand landmarks. We contrast our methods with leading state-of-the-art techniques in body posture and hand motion generation. For this purpose, we replicate body and hand generation methods independently. As detailed in Table 2, with three-frame seed poses, our method exceeds previous state-of-the-art algorithms.

Table 2. Multi-metric Performance Comparison with State-of-the-Art Methods (↓ indicates lower is better, ↑ indicates higher is better, bold indicates the best value).

	FGD ↓	FID ↓	Diversity ↑	MSE ↓	LVD ↓
Baseline	13.080	6.941	8.3145	1.442	9.317
+VQVAE	9.787	6.673	10.624	1.619	9.473
+4 VQVAE	7.397	6.698	12.544	1.243	8.938
FACT	6.673	**6.371**	12.954	1.203	8.998
+Masked Hints	5.423	6.794	13.057	1.180	9.015
PTDM (ours)	**5.214**	6.641	**13.213**	**1.091**	**8.265**

We employ FGD [49] to evaluate the realism of body posture, measuring Diversity [51] by calculating the mean L1 distance across multiple body posture segments. For facial expressions, we calculate the vertex MSE [52] to assess the positional distance and vertex LVD [53] between ground-truth and generated facial vertices. This approach maintains the authenticity of body movements while enhancing the facial details to resemble real-world scenarios more closely.

With the availability of 3D ground-truth scans, we can compute the error of rendered textures based on camera-space pixels. Under the multiview evaluation protocol for high-resolution images, as detailed in Table 3, SMPLitex surpasses the state-of-the-art method [19] in both Structural Similarity Index Measure (SSIM) [54] and Learned Perceptual Image Patch Similarity (LPIPS) [55] metrics.

Table 3. SSIM and LPIPS Performance Comparison Across Methods (↑ indicates higher is better, ↓ indicates lower is better, bold indicates the best value).

	SSIM ↑	LPIPS ↓
CMR [56]	0.7142	0.1275
HPBTT [20]	0.7420	0.1168
RSTG [16]	0.6735	0.1778
TexGlo [19]	0.6658	0.1776
SMPLitex [57]	0.8648	0.0695
ATGT3D (ours)	**0.9173**	**0.0215**

We evaluated our method both qualitatively and quantitatively on the AMASS dataset and five publicly accessible datasets from BEAT2 [32], demonstrating performance that matches leading texture estimation methods. Compared to existing methods, SMPLitex [57] accommodates both low-resolution and high-resolution images and maintains robustness across multiview consistency metrics. An upward arrow in the SSIM section indicates that a higher SSIM value signifies better similarity between the reconstructed image and the ground truth (the higher the value, the better the similarity). In the LPIPS section, a downward arrow indicates that a lower LPIPS value represents better perceptual similarity (the lower the value, the less distortion).Bold values highlight the best performance in each metric.

The PTDM proposed in this paper significantly enhances human motion realism. Limitations in simulating garment movement also emerge, particularly with light and flowing fabrics such as silk. The discrepancy between the movement trajectories of the clothing and the human body results in blurred textures and body details. Future research will focus on integrating physical simulations to bolster model robustness across diverse materials and complex motions or incorporating additional datasets to improve adaptability for intricate garment dynamics.

6. Conclusions

We propose a comprehensive texture mapping method, animating 3D virtual human texture tracking (ATGT3D), which introduces the EDM to estimate the complete 3D human

texture from a single image using a seed image of human texture. Additionally, we present the PTDM, designed to achieve precise texture tracking for the animation of 3D virtual humans based on SMPL-X body parameters. The results with our modules exceed those of current methods based on Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs). Despite the high quality of our results, ATGT3D has limitations, particularly when the clothing in the input image is in motion, such as being blown by the wind away from the body, which can degrade ATGT3D's sampling conditions, leading to mismatches between the texture and the body. Our future work will aim to naturally reconstruct any clothing onto the 3D human model.

Author Contributions: Writing—original draft, F.C.; Writing—review & editing, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available in [BEAT2, AMASS, HUMBI] at [https://doi.org/10.48550/arXiv.2401.00374 (accessed on 10 February 2024), https://doi.org/10.48550/arXiv.1904.03278 (accessed on 15 April 2024), https://doi.org/10.48550/arXiv.1812.00281 (accessed on 11 March 2022)], reference number [30,35,36]. These data were derived from the following resources available in the public domain: [https://huggingface.co/datasets/H-Liu1997/BEAT2/viewer (accessed on 10 February 2024), https://amass.is.tue.mpg.de/ (accessed on 15 April 2024), https://humbi-data.net/ (accessed on 11 March 2022)].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fu, J.; Li, S.; Jiang, Y.; Lin, K.Y.; Qian, C.; Loy, C.C.; Wu, W.; Liu, Z. Stylegan-human: A data-centric odyssey of human generation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 1–19.
2. Grigorev, A.; Iskakov, K.; Ianina, A.; Bashirov, R.; Zakharkin, I.; Vakhitov, A.; Lempitsky, V. Stylepeople: A generative model of fullbody human avatars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5151–5160.
3. Lewis, K.M.; Varadharajan, S.; Kemelmacher-Shlizerman, I. Tryongan: Body-aware try-on via layered interpolation. *ACM Trans. Graph.* **2021**, *40*, 1–10. [CrossRef]
4. Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.Y.; Lian, Z. Controllable person image synthesis with attribute-decomposed gan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5084–5093.
5. Pumarola, A.; Agudo, A.; Sanfeliu, A.; Moreno-Noguer, F. Unsupervised person image synthesis in arbitrary poses. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8620–8628.
6. Sarkar, K.; Golyanik, V.; Liu, L.; Theobalt, C. Style and pose control for image synthesis of humans from a single monocular view. *arXiv* **2021**, arXiv:2102.11263.
7. Sarkar, K.; Liu, L.; Golyanik, V.; Theobalt, C. Humangan: A generative model of human images. In Proceedings of the 2021 International Conference on 3D Vision (3DV), London, UK, 1–3 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 258–267.
8. Vahdat, A.; Kreis, K. *Improving Diffusion Models as an Alternative to GANs, Part 1. NVIDIA Technical Blog*; NVIDIA Developer: Santa Clara, CA, USA, 2022.
9. Guo, C.; Zuo, X.; Wang, S.; Cheng, L. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 580–597.
10. Ebert, D. *Texturing & Modeling: A Procedural Approach*; Morgan Kaufman: San Fransisco, CA, USA, 2002.
11. Jiang, W.; Yi, K.M.; Samei, G.; Tuzel, O.; Ranjan, A. Neuman: Neural human radiance field from a single video. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 402–418.
12. Noguchi, A.; Sun, X.; Lin, S.; Harada, T. Neural articulated radiance field. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 5762–5772.
13. Peng, S.; Dong, J.; Wang, Q.; Zhang, S.; Shuai, Q.; Zhou, X.; Bao, H. Animatable neural radiance fields for modeling dynamic human bodies. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 14314–14323.
14. Prokudin, S.; Black, M.J.; Romero, J. Smplpix: Neural avatars from 3d human models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2021; pp. 1810–1819.

15. Weng, C.Y.; Curless, B.; Srinivasan, P.P.; Barron, J.T.; Kemelmacher-Shlizerman, I. Humannerf: Free-viewpoint rendering of moving people from monocular video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 16210–16220.
16. Wang, J.; Zhong, Y.; Li, Y.; Zhang, C.; Wei, Y. Re-identification supervised texture generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 11846–11856.
17. Jiang, Y.; Yang, S.; Qiu, H.; Wu, W.; Loy, C.C.; Liu, Z. Text2human: Text-driven controllable human image generation. *ACM Trans. Graph. (TOG)* **2022**, *41*, 1–11. [CrossRef]
18. Neverova, N.; Guler, R.A.; Kokkinos, I. Dense pose transfer. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 123–138.
19. Xu, X.; Loy, C.C. 3D human texture estimation from a single image with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 13849–13858.
20. Zhao, F.; Liao, S.; Zhang, K.; Shao, L. Human parsing based texture transfer from single image to 3D human via cross-view consistency. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 14326–14337.
21. Lazova, V.; Insafutdinov, E.; Pons-Moll, G. 360-degree textures of people in clothing from a single image. In Proceedings of the 2019 International Conference on 3D Vision (3DV), Quebec City, QC, Canada, 16–19 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 643–653.
22. Alldieck, T.; Zanfir, M.; Sminchisescu, C. Photorealistic monocular 3d reconstruction of humans wearing clothing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 1506–1515.
23. He, T.; Xu, Y.; Saito, S.; Soatto, S.; Tung, T. Arch++: Animation-ready clothed human reconstruction revisited. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 11046–11056.
24. Li, Z.; Zheng, Z.; Zhang, H.; Ji, C.; Liu, Y. Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 322–341.
25. Natsume, R.; Saito, S.; Huang, Z.; Chen, W.; Ma, C.; Li, H.; Morishima, S. Siclope: Silhouette-based clothed people. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4480–4490.
26. Zheng, Z.; Yu, T.; Liu, Y.; Dai, Q. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3170–3184. [CrossRef] [PubMed]
27. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* **2015**, *34*, 248:1–248:16. [CrossRef]
28. Kurita, T. Principal component analysis (PCA). In *Computer Vision: A Reference Guide*; Springer: Cham, Switzerland, 2019; pp. 1–4.
29. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.A.; Tzionas, D.; Black, M.J. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
30. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv* **2021**, arXiv:2112.10752. Available online: http://arxiv.org/abs/2112.10752 (accessed on 20 December 2021).
31. Grigorev, A.; Sevastopolsky, A.; Vakhitov, A.; Lempitsky, V. Coordinate-based texture inpainting for pose-guided human image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 12135–12144.
32. Liu, H.; Zhu, Z.; Becherini, G.; Peng, Y.; Su, M.; Zhou, Y.; Iwamoto, N.; Zheng, B.; Black, M.J. Emage: Towards unified holistic co-speech gesture generation via masked audio gesture modeling. *arXiv* **2023**, arXiv:2401.00374.
33. Cheong, S.Y.; Mustafa, A.; Gilbert, A. Kpe: Keypoint pose encoding for transformer-based image generation. *arXiv* **2022**, arXiv:2203.04907.
34. Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; Liu, Z. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv* **2022**, arXiv:2205.08535. [CrossRef]
35. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML), PMLR, Virtual Event, 18–24 July 2021; pp. 8748–8763.
36. Loper, M.; Mahmood, N.; Black, M.J. MoSh: Motion and shape capture from sparse markers. *ACM Trans. Graph.* **2014**, *33*, 220. [CrossRef]
37. Mahmood, N.; Ghorbani, N.; Troje, N.F.; Pons-Moll, G.; Black, M.J. AMASS: Archive of motion capture as surface shapes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5442–5451.
38. Yu, Z.; Yoon, J.S.; Lee, I.K.; Venkatesh, P.; Park, J.; Yu, J.; Park, H.S. Humbi: A large multiview dataset of human body expressions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2990–3000.
39. Krebs, F.; Meixner, A.; Patzer, I.; Asfour, T. The KIT Bimanual Manipulation Dataset. In Proceedings of the IEEE/RAS International Conference on Humanoid Robots (Humanoids), Munich, Germany, 18–20 July 2021; pp. 499–506.

40. Firmani, F.; Park, E.J. A framework for the analysis and synthesis of 3D dynamic human gait. *Robotica* **2012**, *30*, 145–157. [CrossRef]
41. Cai, Y.; Wang, Y.; Zhu, Y.; Cham, T.J.; Cai, J.; Yuan, J.; Liu, J.; Zheng, C.; Yan, S.; Ding, H.; et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 11645–11655.
42. Ghorbani, S.; Mahdaviani, K.; Thaler, A.; Kording, K.; Cook, D.J.; Blohm, G.; Troje, N.F. MoVi: A large multi-purpose human motion and video dataset. *PLoS ONE* **2021**, *16*, e0253157. [CrossRef] [PubMed]
43. Mandery, C.; Terlemez, O.; Do, M.; Vahrenkamp, N.; Asfour, T. The KIT Whole-Body Human Motion Database. In Proceedings of the International Conference on Advanced Robotics (ICAR), Istanbul, Turkey, 27–31 July 2015; pp. 329–336.
44. Mandery, C.; Terlemez, O.; Do, M.; Vahrenkamp, N.; Asfour, T. Unifying Representations and Large-Scale Whole-Body Motion Databases for Studying Human Motion. *IEEE Trans. Robot.* **2016**, *32*, 796–809. [CrossRef]
45. Guler, R.A.; Natalia Neverova, I.K. DensePose: Dense Human Pose Estimation in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
46. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. Available online: https://github.com/facebookresearch/detectron2 (accessed on 1 November 2019).
47. Popescu, M.C.; Balas, V.E.; Perescu-Popescu, L.; Mastorakis, N. Multilayer perceptron and neural networks. *WSEAS Trans. Circuits Syst.* **2009**, *8*, 579–588.
48. Kim, J.; Cho, H.; Kim, J.; Tiruneh, Y.Y.; Baek, S. Sddgr: Stable diffusion-based deep generative replay for class incremental object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 28772–28781.
49. Yoon, Y.; Cha, B.; Lee, J.H.; Jang, M.; Lee, J.; Kim, J.; Lee, G. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph. (TOG)* **2020**, *39*, 1–16. [CrossRef]
50. Soloveitchik, M.; Diskin, T.; Morin, E.; Wiesel, A. Conditional frechet inception distance. *arXiv* **2021**, arXiv:2103.11521.
51. Li, J.; Kang, D.; Pei, W.; Zhe, X.; Zhang, Y.; He, Z.; Bao, L. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 11293–11302.
52. Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; Wong, T.T. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 12780–12790.
53. Yi, H.; Liang, H.; Liu, Y.; Cao, Q.; Wen, Y.; Bolkart, T.; Tao, D.; Black, M.J. Generating holistic 3d human motion from speech. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 469–480.
54. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
55. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595.
56. Kanazawa, A.; Tulsiani, S.; Efros, A.A.; Malik, J. Learning category-specific mesh reconstruction from image collections. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 371–386.
57. Casas, D.; Comino-Trinidad, M. SMPLitex: A Generative Model and Dataset for 3D Human Texture Estimation from Single Image. In Proceedings of the British Machine Vision Conference (BMVC), Aberdeen, UK, 20–24 November 2023.