International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-17

Article

Engineering Animated Characters for Cross-Cultural Markets: AI-Assisted Parametric Modeling and Cultural Identity Adaptation

Jihong Huang [1], Shafilla Subri [1,*], Faryna Mohd Khalis [2] and Rui Li [2]

[1] Universiti Teknologi MARA, Merbok sungai petani kedah, 08000, Malaysia
[2] Universiti Teknologi MARA, Jalan Ilmu 1/1, 40450 Shah Alam, Selangor, Malaysia
* Correspondence author: shafilla@163.com

Abstract: 3D animation character modeling is not only time-consuming and labor-intensive, but also has high requirements on the modeler's professional skills. For this reason, this paper proposes a parametric model regression algorithm based on a two-branch structure from the problems existing in animation character modeling. The algorithm obtains the corresponding shape parameters and pose parameters as a priori information by matching the parametric model, so that the implicit function can be effectively sampled and mapped to the point set in the reconstruction. At the same time, the shape parameters and pose parameters of RaBit are regressed separately, and multiple rounds of regression are performed on the pose parameters using the geometric features as reference information in order to realize the fine-tuning of the pose parameters. On the three datasets of Render People, THUman 2.0 and Deep Human, the PSNR values in terms of reconstruction accuracy are 30.83, 29.93 and 31.08, and in terms of rendering efficiency are 64, 48 and 73 FPS, respectively, which achieve a good balance between reconstruction accuracy and rendering efficiency. Finally, the value of animation characters in spreading and transmitting national culture and strengthening cultural identity is discussed, and a cultural identity suitable for the development of domestic animation is constructed, which provides a new direction for animation character reconstruction.

Keywords: animation character modeling; two-branch structure; parametric model; RaBit; cultural identity

1. Introduction

Animated characters are characters that are presented in the form of cartoons, animation, and hand-drawings. These animated characters may be either based on real-life abstract concepts or originated from the imagination and fiction of artists [1]. Their appearance design is usually imaginative and creative, with distinctive features for easy recognition. Animated characters play an important role in entertainment works, which can trigger the audience's emotional engagement and empathy, and are often used to convey the author's views, values or educational significance [2-3]. In addition to entertainment, animated characters are widely used in advertising, education, social media and other aspects, and have a very wide audience. Characters with high popularity can promote the development of related industries through the sale of dolls, stationery, clothing and other derivative products of their images [4]. Currently, with the rapid development of animation, games, movies and other fields, the demand for high-quality, personalized animation character models is growing [5]. These models play a crucial role in the works and can provide a richer and deeper experience for the audience or players. Although manual modeling techniques are well established in existing industrial processes, traditional manual modeling still faces numerous challenges, such as high technical difficulty, high entry barrier, and low modeling efficiency [6]. Manual modeling not only requires the

modeler, who has rich professional knowledge, but also needs to be proficient in professional modeling software such as C4D, Blender, etc., as well as have a certain art foundation, which greatly enhance the technical threshold. Therefore, there is an urgent need for a more efficient and flexible parametric modeling technology to meet the growing market demand. As a result, animation character modeling technology based on artificial intelligence came into being, bringing revolutionary changes to the field of animation modeling [7].

Artificial Intelligence (AI) technologies represented by machine learning and deep learning have brought significant changes to the engineered animation character modeling industry, both improving the modeling efficiency and optimizing the modeling accuracy [8]. In the past, constructing a parametric animation character model was extremely dependent on tedious manual sculpting and rigorous mathematical derivation, while the focus of academic research has now shifted to how to use AI to automatically refine parametric laws from massive data. Luo, Z et al. proposed a large-scale dataset 3DBiCar and corresponding parametric model RaBit for 3D bipedal cartoon characters. The dataset contains 1500 high-quality 3D texture models handcrafted by professional artists with consistent topology, and RaBit is designed with a SMPL-like linear hybrid shape model with a StyleGAN-based neural UV texture generator that can represent shape, pose and texture simultaneously [9]. Yu, S et al. introduced machine vision and artificial intelligence technology into animation design based on CAD parametric design methodology, i.e., fast design modification through parameter constraints and parameter value adjustments, and realized automatic identification, extraction and generative scene design of key elements in animated scenes [10]. Experiments by Yang, Z and Yin, Z on two types of physical character control tasks, character morphology optimization and Deep Mimic hyper-parameter tuning, show that the algorithm significantly outperforms existing hyper-parameter optimization methods applicable to physical character animation, and strengthens the ability of parametric character models to integrate the three aspects of geometric representation, motion control and training efficiency [11]. Eckert et al. proposed Scalar Flow, a framework that contains two core components, a novel estimation algorithm for invisible inflow regions, and an efficient optimization scheme subject to simulation constraints for capturing real fluids, which provides high-fidelity training data and reconstruction methods for simulation-driven generation of complex smoke animations [12]. Lin, J et al. proposed Motion-X, a large-scale 3D embodied full-body motion dataset, which contains 15.6 million precise 3D full-body pose annotations and covers 811,000 action sequences. This initiative has driven the field of character animation to shift from "motion reproduction" to a "semantically-driven" generation paradigm [13].

After obtaining a high-precision static model, how to make it move without wearing out the mold has become a hot topic in academic research. For a long time, the industry has relied on the “Linear Blending Skinning (LBS)” technique, which often makes the characters look like “plastic toys” in animation, and is prone to volume loss or joint distortion. To address this pain point, Zhang, H et al. proposed PhysRig, which a physically based framework for binding micromountable skins to bones by embedding a rigid skeleton into a volumetric representation (e.g., tetrahedral mesh) and modeling it as a deformable soft-tissue structure driven by an animated skeleton, thus overcoming the traditional linear hybrid skin's volumetric losses, unnatural deformations, and inability to model soft tissues, hairs , elastic materials such as flexible appendages and other limitations [14]. Based on PhysRig, Zhang, H and Luo, J proposed a unified generative framework-RigMo-that jointly learns bindings and motions directly from raw mesh sequences without any manually provided binding annotations [15]. Tan, S et al. proposed a generalized LDM-based animation framework, Animate-X, which introduces gesture indicators to capture comprehensive motion patterns from the driving video both implicitly and explicitly, supporting a wide range of character types including anthropomorphic characters [16]. Gui, Z et al. addressed the long-standing "extreme appearance diversity" and "long-tail data distribution" challenges in animation character recognition by constructing a cross-modal (visual + audio) character knowledge base. This provided a systematic solution for the intelligent understanding of animation characters and the generation of accessible content [17]. Gao, J et al. proposed Character Shot, which achieves end-to-end generation from single-image reference and gesture sequences to high-quality 4D character animation through the coupling of 2D generative models with 4D Gaussian splashes [18]. Zhang, N et al. intelligently analyzed and optimized animated scenes and character portrayals by integrating multimodal emotion recognition techniques using advanced convolutional neural networks with long and short-term memory models [19].

With the explosion of big models and generative AI (AIGC), parametric modeling of characters is gradually moving away from cumbersome slider adjustments to intuitive semantic control.Bai, Z et al. proposed a comprehensive solution for automated generation of avatar facial animation, which redirects the facial expressions of the input face image to the avatar face by estimating the blend shape coefficients, and supports different appearances with the blend shape topology of character animation

generation [20]. Tang, Y et al. pointed out that the key path for generative AI to reshape animation production lies in upgrading the tool from “assisted portrayal” to “intelligent co-creation”, so that animation production can be shifted from labor-intensive to creativity-driven [21]. Zarif and others have demonstrated the contributions of generative AI in animation production. Firstly, it serves as a creative engine to expand the breadth and efficiency of visual conception. Secondly, a structured content controllability and quality evaluation system needs to be established to bridge the gap between "machine generation" and "professional deployment" [22]. Lungu-Stan, V and Mocanu, I optimize traditional animation workflows by predicting the next pose based on a sequence of historical poses, while combining historical poses with current noise poses to predict clean poses and process motion capture data for noise reduction [23]. Qiu, S constructed a generative AI pipeline from concept generation (language/image modeling) to controllable animation output, focusing on solving the two bottlenecks of “poor style consistency” and “uncontrollable randomness” of traditional generative modeling in game animation, providing a new path to rapid prototyping and iterative production for 2D It provides a new path for rapid prototyping and iterative production of 2D game character animation [24]. Izani, M et al. pointed out that the “motion defect” of AI-generated animation is rooted in the lack of spatial-temporal dynamic modeling ability, and it is difficult to capture the temporal tuning of motion, the sense of weight, and the intrinsic rhythm of anticipation-response with purely visual diffusion or linguistic modeling, so it is necessary to develop a generative framework that integrates the physical perception and the semantics of motion [25]. Clocchiatti, A et al. proposed a higher-order pipeline for generating 2D characters and animations from scripts, combining a latent diffusion model with a large-scale language model, which uses ChatGPT to generate character descriptions from scripts, which in turn generates customized character images using Stable Diffusion, and animates them according to the character's movements in different scenes [26].

In this study, a regression algorithm for parameterized models of animated characters based on a two-branch structure is proposed around the problems of diversity of morphology, exaggeration and complexity of poses that occur in animated character modeling. The algorithm regresses the morphological parameters and pose parameters of the parameterized model through two branches respectively. At the same time, the corresponding shape parameters and pose parameters are obtained by matching the parameterized model to be used as a priori information, which enables the implicit function to carry out effective sampling and point set mapping in the reconstruction. In the two-branch structure, the morphological branch utilizes the line information in the image to help the regressor regress the morphological parameters, while the pose branch obtains more accurate pose parameters through multiple fine-tuning with the help of geometric features.

2. Methodology

2.1. Parameterized models

The 3D animation bipedal character parametric model RaBit is a recently proposed parametric modeling method for cartoon characters, which optimizes and improves the human body parametric model SMPL for animated characters, making the parametric model more suitable for cartoon characters, and retains the advantages of SMPL, which can be directly used in games, animation and other practical application scenarios. The differences between SMPL and RaBit are shown in Table 1. RaBit, like SMPL, splits the parameterized model parameters into two parts: morphology and pose. Its morphological parameters β are similar to the body shape parameters proposed in SMPL, which are also obtained by extracting the model mesh from the 3D character dataset for principal component analysis. In RaBit, the morphological parameters are vectors of length 100. This coefficient is used to control the character's height, weight, species, size, and other features related to body shape.

Table 1. Comparison of SMPL and RaBit Attributes

Parameterized model properties	SMPL	RaBit
Vertex quantity	6850	39225
Number of nodes	21	21
Number of mesh surfaces	13552	75382
Number of morphological parameters	10	100
Number of attitude parameters	72	72

The process of modeling the parametric model by morphological parameters is as follows:

$$\mathrm{M}_{t} = \overline{\mathrm{M}_{t}} + \sum_{i=1}^{m} (\beta_i \cdot s_i) \tag{1}$$

where M_{t} denotes the T-Pose parameterized model modeled using attitude parameters. $\overline{\mathrm{M}_{t}}$ denotes the initial T-Pose parameterized model. $m = 100$ denotes the length of the morphology parameter is 100. $s_i \in \mathbb{R}^{3\times N}$ denotes the vertex offsets controlled by different morphology parameters.

The RaBit pose parameter $\theta = [w_0^T, \cdots, w_K^T]^T \in \mathbb{R}^{3\times K}$ is kept consistent with that in SMPL to represent the pose of the cartoon character. Where w_0^T denotes the rotation of the root node of the parameterized model, $K = 24$ denotes a total of 23 joints and 1 root node, and the parameter $w_k \in \mathbb{R}^3$ for each joint denotes the axial angle representation of the k th joint with respect to the parent bone in its skeletal structure, which can be converted into the corresponding rotation by the Rodriguez formula. matrix, which controls the pose of RaBit.RaBit utilizes a vertex-based linear hybrid skin control parameterized model for pose change, a process under which this pose change occurs:

$$v_i' = \sum_{k=1}^{K} \left[W_{k,i} \cdot G_k(\theta, J) \cdot v_i \right] \tag{2}$$

where v_i' , v_i denote the vertex positions after and before the pose change, respectively. $W_{k,i}$ denotes the weight of the current vertex affected by the rotation matrix of different joints. The $G_k(\theta, J)$ then denotes the rotation matrix of the joint point k computed from the pose parameter θ and the joint point position J .

2.2. Extraction of a priori information on model parameters

2.2.1. Static model generation

In order to construct a highly realistic 3D animated character model, the key image information contained in the input image must be fully explored. This includes information about the pose data of the animated character, information about the details of the contour edges, and information about the texture features of the clothing. By integrating these elements, a 3D animated character model with a sense of realism in 3D space can be effectively reproduced.

In order to obtain the image information of the main body of the animated character image, it is first necessary to perform segmentation processing on the image, with the aim of generating a mask map corresponding to it. This process involves accurate identification and segmentation of the animated character in the image. Then, a mask map is created in which the character region is labeled as one category and the background as another. This binarized mask map plays an important role in further image processing and analysis. In this paper, DeepLabv3+ algorithm is used to perform preprocessing operations on the input image, including image segmentation and mask generation.DeepLabv3+ is the latest iteration of semantic segmentation algorithm of the DeepLab family of algorithms.For the above mentioned problem of balancing spatial location and semantic information, DeepLab family of algorithms employs a Dilated FCN structure, which uses the null convolution to expand the sensory field and increase the effective size of the standard convolutional kernel by inserting voids into it to capture a wider range of contextual information while keeping the computational efficiency and the number of parameters unchanged.

Subsequently, the masked image obtained from the preprocessing network is fed into the PIFu network along with the original image.

A functional representation of the figure in implicit space is generated.The PIFu network utilizes a pixel-to-pixel implicit function representation method to associate each pixel in the 2D image with the corresponding 3D human body information.The framework takes a single 2D photo of the human body and its mask image as input, and is capable of generating a 3D model of the human body and its texture information (RGB information is specifically referred to in the method).The workflow of the PIFu network is illustrated in Fig. 1 As shown in the training phase, for the 2D image and its mask map, pixel-level feature extraction is performed using an encoder consisting of a convolutional neural network (CNN), which generates a high-dimensional feature vector (deep feature) for each pixel. Spatial point sampling is performed for the 3D animated character model, which is used to generate the real model used for judgment in the training phase.

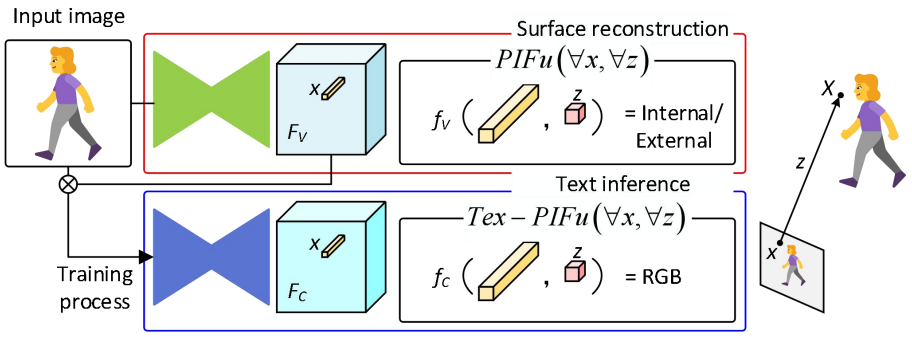

Figure 1. Flowchart of the PIFu network workflow

Uniform and Gaussian sampling 1:16 offset sampling is used in this topic. From the above known information, the loss function is constructed as shown below:

$$L_v = \frac{1}{n}\sum_{i=1}^{n}\left|f_v(F_V(X_i), z(X_i)) - f_v^*(X_i)\right|^2 \tag{3}$$

where n is the number of sampling points in the 3D space of the model, the corresponding $f_v^*(X_i)$ is the Ground Truth point set, and f_v is the implicit expression corresponding to the image sampling points and their depth features obtained from the multilayer perceptron fitting. After completing the above training, the inference stage can utilize the trained implicit function to output 3D occupied field information for the input 2D image, and obtain the corresponding animated character mesh model through the equivalent surface extraction algorithm.

2.2.2. Parameter Matching Function Design

The model parameter a priori information defined in this paper refers to the morphological parameters β and attitude parameters θ of the parameterized model obtained from fitting the generated static model. To ensure the consistency and generality of the parameters, the parameterized model used for fitting and matching is the SMPL model. Therefore, the content of this section is to fit the matching parameterized model according to the function expression of the existing static model and solve its morphology parameter β and attitude parameter θ.

In this paper, the theoretical support is obtained from the idea of transforming model fitting into parameter optimization in the SMPLify-X reconstruction algorithm, which transforms the solved problem into the distance between the existing static model and the parameterized model, minimizes the problem, and further improves the loss function from the two perspectives of the unnatural attitude penalty and the collision penetration penalty, and the complete flow is shown in Figure 2.

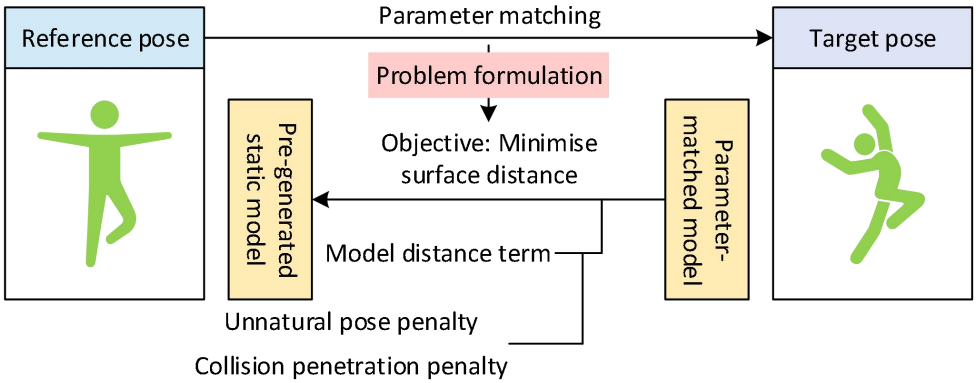

Figure 2. Prior information parameter matching process

Regarding the definition of the SMPL model, it can be seen that the solution model M is to solve for the fitted deformed morphology parameter β and attitude parameter θ. Since the number of vertices and the mesh structure are different between M and P and there is no point-to-point correspondence, the distance optimization term proposed in this section is calculated using the chamfer distance. Chamfer

distance is a metric used to measure the difference between two point sets, which is calculated as the average of the distances from a point in two point sets to the nearest point in the other point set, so the chamfer distance is insensitive to the order of the points in the point sets and does not rely on the exact correspondence between two point sets, which meets the requirements for the calculation of the distance optimization term in this section. Let A, B be two point sets whose chamfer distance $Dchamfer(A,\ B)$ is calculated as shown below:

$$\begin{aligned} D_{chamfer}(A,B) = &\frac{1}{|A|}\sum_{a\in A}\min_{b\in B}\|a-b\|_2^2 \\ &+\frac{1}{|B|}\sum_{b\in B}\min_{a\in A}\|b-a\|_2^2 \end{aligned} \tag{4}$$

where $\|a-b\|_2$ is the Euclidean distance between point a and point b.

According to the above equation in the same way can be chamfer distance measure between M and P distance minimization problem of the objective function is:

$$E_{mesh}(\beta,\theta) = E_{MP}(\beta,\theta) + E_{PM}(\beta,\theta) \tag{5}$$

$$E_{MP}(\beta,\theta) = \sum_{x\in M(\beta,\theta)}\min_{y\in P}\|x-y\|_2^2 \tag{6}$$

$$E_{PM}(\beta,\theta) = \sum_{y\in P(\beta,\theta)}\min_{x\in M}\|y-x\|_2^2 \tag{7}$$

where E_{MP} denotes the sum of the minimum distances from the vertices in the parameterized model M to the static model P; E_{PM} denotes the sum of the minimum distances from the vertices in the static model P to the parameterized model M.

For the unnatural posture penalty term, which aims to ensure that the model fitting results are biomechanically feasible and avoid producing unrealistic or physiologically impossible postures, corresponding penalty values need to be given for unreasonable bending angles and bending directions of the joints. In this paper, we fit the SMPL model to the labeled poses in the CMU dataset, probabilistically modeling possible human postures and constructing posture prior terms. In this model, each Gaussian component represents a subregion in the human pose space, and the mixing coefficients represent the probabilistic weights of the subregion. During the optimization process, a probabilistic framework that conforms to the statistical distribution of human body postures is obtained, and its final Gaussian mixing formula and the derivation process are shown below:

$$\begin{aligned} E_\theta(\theta) &\equiv -\log\left(g_j N\left(\theta;\mu_{\theta,j},\sum_{\theta,j}\right)\right) \\ &\approx -\log\left(\max_j\left(cg_j N\left(\theta;\mu_{\theta,j},\sum_{\theta,j}\right)\right)\right) \\ &= \min_j\left(-\log\left(cg_j N\left(\theta;\mu_{\theta,j},\sum_{\theta,j}\right)\right)\right) \end{aligned} \tag{8}$$

where g_j is the weight of the $N=8$ Gaussian mixture model, and the original function is computed by taking the negative of the logarithm of the weighted sum of multiple Gaussian distributions (each consisting of the mean $\mu_{\theta,j}$ and covariance $\sum 1_{\theta,j}$ is defined); to simplify the computation, the original function is approximated as taking the negative of the logarithm of the largest term, where c is a constant factor used to ensure numerical stability; and then transforming the notation converts the approximation formula into a minimization form, i.e., looking for the value of j that minimizes the original function, and finding the Gaussian distribution that best describes the current attitude θ.

Combining the above formula information, the pose penalty terms defined in this section for penalizing the fitting process are shown below:

$$E_{prior}(\theta) = \sum_i exp(\theta_i) + E_\theta(\theta) \tag{9}$$

where i denotes the joint ordinal number, θ is the rotation angle of the joint, and the first part amplifies the penalty term for unnatural postures in the form of an exponential function, where $\theta_i = 0$ when the joint is not currently curved, and is not heavily penalized for negative curvatures $\theta_i < 0$, and

for positive curvatures $\theta_i > 0$, the exponential function increases the penalty value exponentially with the positive rotation angle.

2.3. Algorithm design

2.3.1. Twin-branch structure

In traditional SMPL-based 3D parameter regression algorithms for mannequins, a single-branch network structure is often used to regress both morphological and pose parameters using a single regressor after extracting image features through a backbone network. Many methods focus on optimizing how to obtain more accurate pose parameter regression on this basis, in order to obtain more accurate pose parameters. These methods have achieved remarkable results in animated character reconstruction, but have neglected the importance of morphological parameters, which is more significant when dealing with animated character models with complex shapes. The main reason is that RaBit's morphological parameters can express more information compared to SMPL, including different kinds of cartoon characters, height, thickness, fatness and other information. The number of parameters is 10 times that of SMPL. And the cartoon character posture often does not conform to the physical law, with exaggerated performance. This makes the traditional single branch, single regression approach may have difficulties in regressing accurate parameterized model parameters.

To address this problem, this algorithm uses a two-branch structure to regress the parameterized model parameters, and its workflow is shown in Figure 3. This design aims to focus on both the morphological parameters and the pose parameters of the parameterized model of the animated character, so as to achieve a more accurate reconstruction of the parameterized model of the animated character. Specifically, this algorithm uses independent ResNet50 on the morphological branch and the pose branch, respectively, as the feature encoder of the corresponding branches. After the input single-view animated character images are feature extracted by the encoders of different branches, the image features for regression of morphological parameters are obtained respectively.

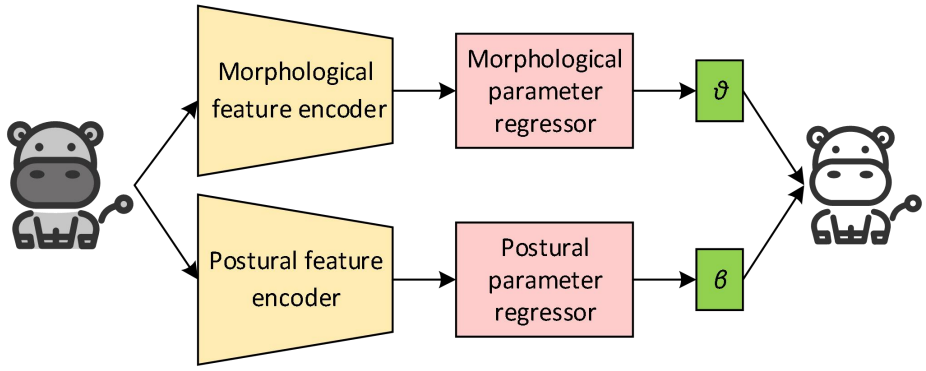

Figure 3. Workflow diagram of the dual-branch structure

2.3.2. Geometric feature fine-tuning module

In addition, the movements of animation characters often contain some purposes of the artist or author, and their movements sometimes have a certain degree of abstraction and exaggeration. In addition, the types of animated characters are very diverse, and different types of animated characters may have different performances when making the same movements. Considering the above problems, it is relatively difficult for traditional SMPL-based methods to obtain accurate pose parameters using single regression. To solve this problem, this paper proposes a geometric feature fine-tuning module (GFF). This module takes the geometric features as reference information, inputs the geometric features together with the picture features into the regressor, and performs multiple rounds of regression on the resulting pose parameters for fine-tuning and optimization to obtain more accurate prediction results.

The workflow of the GFF module is shown in Fig. 4. After the model mesh M is reconstructed using the pose parameters θ of the T-Pose and the predicted morphology parameters β, M is downsampled to reduce the number of vertices, and the number of vertices is lowered from 37,152 to 605 to obtain the model mesh M' with the reduced number of vertices. Subsequently, all the vertices of M' and the joints of the parameterized model are combined together and input into the geometric feature extractor to extract the geometric features contained in the vertices and joints, and the obtained

geometric features and picture features are fused together and input into the pose parameter regressor to obtain the new pose parameters. After obtaining the new pose parameters, we continue to reconstruct the parametric model using the predicted morphological parameters and the new pose parameters, and repeat the fine-tuning process of downsampling the model mesh vertices, extracting the geometric features, combining the picture features, regressing the pose parameters, and parametrically modeling the model. The fine-tuning process was repeated three times to obtain the final attitude parameters.

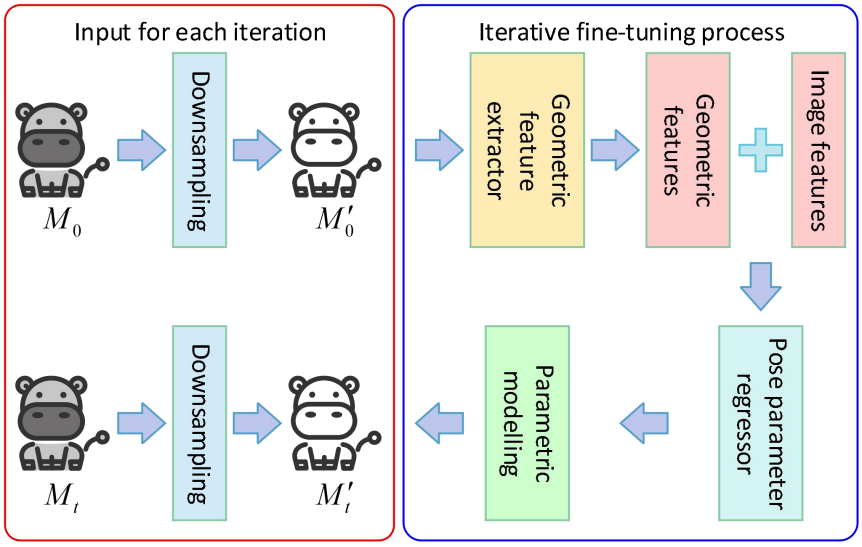

Figure 4. Workflow of the geometric feature fine-tuning module

2.3.3. Loss function design

This study draws on the loss function strategy employed in traditional parametric model regression algorithms in the algorithm design. In traditional parametric model reconstruction algorithms, two types of loss functions are usually employed: a 2D loss function and a 3D loss function. However, for the specific domain of animated characters, whose movements tend to be more exaggerated, and for which the 3D model meshes used for supervision are not acquired by scanning, but are handmade by the artist. This results in large differences between the 3D model and the picture, and makes it difficult to find a plausible projection method that aligns the 3D keypoints with the 2D keypoints of the picture.

Therefore, in this paper, when supervising the neural network, we abandon the 2D loss function and use the 3D loss function to supervise the pictures. The loss functions used in the algorithm of this paper are as follows:

(1) 3D vertex loss

This loss function supervises the neural network by calculating the L1 loss between the model mesh vertices of the predicted parametric model and the model mesh vertices of the real parametric model. Compared to supervising the morphological and pose parameters separately, the vertex-based supervision has the advantage of constraining both pose and morphological parameters, and avoids the jitter and dichotomy problems in parametric supervision. The loss function is formulated as follows:

$$\mathrm{L}_{vertex} = \frac{1}{|V|}\sum_{v\in V}\left|S_v(\beta_v,\theta_v,T_v) - S_v(\hat{\beta}_v,\hat{\theta}_v,\hat{T}_v)\right| \tag{10}$$

where V refers to the vertices of the model mesh, $|V|$ denotes the number of vertices, and $\beta_v,\theta_v,T_v,\hat{\beta}_v,\hat{\theta}_v,\hat{T}_v$ denote the predicted morphology parameter, the attitude parameter, the displacement, and the true values of morphology parameters, attitude parameters, and offsets.

(2) 3D key point loss

This loss function enhances the accuracy of the network for parameter prediction by calculating the mean square error loss between the 3D keypoints of the parameterized model obtained from prediction and the 3D keypoints of the real cartoon model. The calculation formula is shown below:

$$\mathrm{L}_{joint} = \frac{1}{K}\sum_{k=1}^{K}(J_k - \hat{J}_k)^2 \tag{11}$$

where $K = 23$, denotes a total of 23 keypoints. The $J_k, \hat{J}_k$ then indicate different keypoints.

(3) Feature extraction loss

This loss function is used to monitor the correctness of the line extraction module for animated characters. Briefly, this loss supervises the network by calculating the binary cross-entropy loss between the predicted line drawings and the real line drawings. This loss is calculated as follows:

$$\mathrm{L}_{line}(W) = \frac{1}{|I|}\sum_{i=1}^{|I|}\left(\sum_{k=1}^{K} l(x_i^k; W) + l(x_i^{fuse}; W)\right) \tag{12}$$

where $K = 5$, denotes a total of 5 inputs with different resolutions. The x_i^k denotes the current pixel point, and x_i^{fuse} denotes the pixel point obtained after up-sampling the outputs of the line drawings of different resolutions to the same resolution and merging them together, and downscaling them to a single channel by convolution. The function l then represents the computation of the binary cross-entropy loss, and W denotes the weights of the different pixel points. $|I|$ then represents the full number of pixel points in the image.

Finally, the loss function L is calculated as:

$$\mathrm{L} = \lambda_1 \mathrm{L}_{vertex} + \lambda_2 \mathrm{L}_{joint} + \lambda_3 \mathrm{L}_{line} \tag{13}$$

where $\lambda_1 = 100$, $\lambda_2 = 100$, and $\lambda_3 = 10$ denote the different weights taken by different loss functions.

3. Results

3.1. Experimental environment and parameter settings

3.1.1. Experimental environment

The experimental environment of this paper is shown in Table 2.

Table 2. Experimental environment setting

Operating system	**Ubuntu18.04**
Programming language	Python
Programming framework	PyTorch 1.8.0
Computing acceleration	CUDA12.0
GPU	Nvidia GeForce RTX 5090
Memory	64 GB
Video memory	32 GB

3.1.2. Data sets

The Render People dataset has often been used extensively in past animated character reconstruction studies. The dataset contains 500 high-resolution photogrammetric scans that provide high-fidelity geometry and extremely realistic texture effects. However, since the Render People dataset is a commercially available product crafted by artists, it is expensive to build large-scale datasets. Therefore, in some studies, the PIFu method used only 442 human models for network training and another 54 models for testing. It is worth noting that these human models are mostly in upright poses, thus limiting the diversity of postures and body types. In order to achieve a greater diversity of body types and postures, researchers need to seek more publicly available data sources. To expand the diversity of datasets, we evaluated on the publicly available THUman 2.0 and Deep Human datasets.

The THUman 2.0 dataset contains 500 high-quality scans of animated characters, with each scan sample captured by a densely arranged array of digital single-lens reflex cameras (DSLRs). These scans utilize multi-view stereo vision techniques combined with professional-grade photogrammetric equipment to provide millimeter-accurate 3D geometric data and texture information. Each scan sample contains high-resolution RGB texture maps that enable very realistic material rendering effects. Specifically, we used a weak perspective camera to generate 36 RGB images at different yaw angles and extracted the parameters of the animated character joints from each image via OpenPose.

The Deep Human dataset contains 2050 sets of high-precision 3D human body scans, and the data

acquisition process uses an array of 60 Intel RealSense D455 cameras working in synchronization with a Faro laser scanner to ensure millimeter-level geometric accuracy of the data and 4K resolution texture mapping. To further validate the performance of our method, the Deep Human dataset was used as an additional test dataset. These high-quality data sources provide richer and more diverse training data for our animated character reconstruction studies, allowing us to train the model and validate its performance over a wider range of poses and body sizes, thus improving the reliability and generalization of the algorithm in real-world applications.

3.2. Experimental Comparison and Analysis

In order to assess the quality of the appearance of the 3D animated character models reconstructed by the method in this paper, test frame poses from the Render People dataset are used to drive the models and render them. The differences between the rendered images and the actual images are quantified by comparing them. A comprehensive quantitative evaluation of the proposed animated character reconstruction methods is compared.

Several representative algorithms are selected as benchmarks for comparison, including the NeRF-based implicit representation character avatar reconstruction algorithms Human NeRF and Instant Avatar, as well as the recent advanced 3DGS-based display representation algorithms of GART and Gaussian Avatar for quantitative evaluation comparison. The data in Table 3 demonstrates the performance of the method proposed in this paper against other baseline methods for the view synthesis task in a dataset of three character sequences. The results show that the method in this paper outperforms all other methods in terms of peak signal-to-noise ratio. Although the Instant Avatar method receives high scores on the perceptual loss metric and structural similarity index, it does not have real-time rendering capabilities comparable to 3DGS-based methods.

Table 3. Quantitative results of different methods

Method	**Render People dataset**		
	PSNR↑	**SSIM↑**	**LPIPS↓**
Human NeRF	24.31	0.9316	0.0236
Instant Avatar	27.83	**0.9734**	**0.0215**
GART	26.31	0.9626	0.0483
Gaussian Avatar	24.48	0.9613	0.0315
Our method	**30.83**	0.9618	0.0478
	THUman 2.0 dataset		
	PSNR↑	SSIM↑	LPIPS↓
Human NeRF	26.28	0.9692	**0.0193**
Instant Avatar	29.06	0.9683	0.0215
GART	29.31	**0.9718**	0.0382
Gaussian Avatar	29.18	0.9622	0.0224
Our method	**29.93**	0.9625	0.0405
	Deep Human dataset		
	PSNR↑	SSIM↑	LPIPS↓
Human NeRF	26.34	0.0901	0.0193
Instant Avatar	28.28	0.9711	**0.0184**
GART	30.14	**0.9808**	0.0372
Gaussian Avatar	28.06	0.9723	0.0275
Our method	**31.08**	0.9746	0.0306

The method in this paper achieves optimal reconstruction accuracy (PSNR) values of 30.83, 29.93, and 31.08 on the three datasets of Render People, THUman 2.0, and Deep Human, respectively, showing good generalization and robustness. This performance improvement is attributed to the structured scene-guided spatial-temporal reconstruction mechanism, which enables the model to stably predict detailed and continuous local neural Gaussian representations even in the face of camera jitter, occlusion, and geometric errors that are common in real data.

In addition to the reconstruction accuracy, this paper also evaluates the performance of the proposed method in terms of rendering efficiency, and the results are shown in Table 4. As can be seen from the table, the GART method exhibits extremely high rendering speed on all datasets, even up to 213 FPS on the THUman 2.0 dataset, which is the most efficient method currently. However, its reconstruction quality has large limitations in dynamic scenes. In contrast, the animated character reconstruction method proposed in this paper achieves a good balance between reconstruction accuracy and rendering

efficiency. Its rendering efficiency on the Render People, THUman 2.0, and Deep Human datasets is 64, 48, and 73 FPS, respectively, which is significantly better than that of the current mainstream NeRF family of methods (e.g., Human NeRF and Instant Avatar, which are all below 10 FPS), and the rendering efficiency is close to that of some 3DGS-based explicit representation and reconstruction methods (e.g., GART). The efficient rendering performance is attributed to the anchor point spatial structure with a controllable localized neural Gaussian generation mechanism proposed in this paper. Specifically, the number of local neural Gaussians required per frame is dynamically regulated by the anchor-point-driven approach, which enables the model to realize realistic rendering with high efficiency and quality.

Table 4. Rendering efficiency comparison results across different dataset

Method	**FPS↑**		
	Render People	**THUman 2.0**	**Deep Human**
Human NeRF	<10	<10	<10
Instant Avatar	<10	<10	<10
GART	151	213	68
Gaussian Avatar	28	25	29
Our method	64	48	73

3.3. Ablation experiments

In this section, multiple ablation experiments are designed to evaluate and validate the modules in the two-branch structural framework. The evaluation metrics include: the PCK (proportion of correct estimation of keypoints) is used as an error metric for joint alignment; the IoU is used as a metric for contour alignment, comparing the degree of overlap between the predicted results and the true labeling, and measuring the degree of overall fit of the character; and the Surface Normal Similarity (hereinafter referred to as N_sim) is used for evaluating the cosine average of the fitting results and the target surface normal. The contour is a measure of the general shape, and the normal is more used to measure the local shape. In addition, in calculating the surface normal similarity, only sparse regions with surface normal confidence greater than 0.9 were selected.

3.3.1. The Role of Contouring and Normal Error

In order to measure the actual role of contours and surface normals in shape cues, 24 animated character models with different body sizes were selected for this experiment, and the number of images for each animated character was five. The experiments were grouped according to whether shape cues were used or not, and were specifically categorized into three groups: contour only, normal only, and contour+normal. All groups used 2D keypoint cues.The threshold of the PCK metric was 0.05.Table 5 demonstrates the specific results of this experiment.From the average results, the contour+normal group was the best in terms of 2D joints (0.550) and surface normal alignment (0.818), with the normal-only group coming in second place. This shows that the estimation of surface normals does bring some improvement to the overall optimized results.

Table 5. The influence of contour or surface normal error on Fitting accuracy

N	Profile			Normal Direction			Profile +Normal Direction		
	PCK	**IoU**	**N_sim**	**PCK**	**IoU**	**N_sim**	**PCK**	**IoU**	**N_sim**
1	0.253	0.695	0.775	0.443	0.743	0.727	0.567	0.824	0.853
2	0.340	0.760	0.640	0.318	0.714	0.701	0.725	0.693	0.865
3	0.609	0.616	0.671	0.662	0.818	0.816	0.461	0.729	0.827
4	0.626	0.663	0.696	0.303	0.776	0.658	0.449	0.686	0.706
5	0.273	0.739	0.713	0.269	0.816	0.788	0.712	0.776	0.844
6	0.455	0.792	0.607	0.367	0.864	0.747	0.659	0.805	0.858
7	0.630	0.689	0.637	0.281	0.720	0.721	0.432	0.718	0.826
8	0.686	0.647	0.703	0.639	0.840	0.749	0.467	0.799	0.701
9	0.660	0.718	0.773	0.444	0.836	0.741	0.604	0.828	0.860
10	0.722	0.642	0.736	0.674	0.858	0.652	0.302	0.638	0.793
11	0.375	0.649	0.723	0.289	0.703	0.761	0.715	0.697	0.787
12	0.448	0.773	0.686	0.670	0.762	0.824	0.509	0.705	0.873
13	0.283	0.661	0.747	0.498	0.846	0.741	0.456	0.770	0.864
14	0.344	0.784	0.671	0.525	0.864	0.814	0.627	0.838	0.840
15	0.292	0.765	0.799	0.403	0.798	0.735	0.624	0.718	0.814
16	0.681	0.693	0.782	0.442	0.794	0.673	0.700	0.828	0.703
17	0.703	0.690	0.716	0.264	0.733	0.805	0.314	0.632	0.873
18	0.417	0.675	0.728	0.258	0.743	0.702	0.477	0.748	0.786
19	0.263	0.677	0.689	0.355	0.777	0.693	0.730	0.693	0.721
20	0.657	0.743	0.660	0.337	0.855	0.736	0.672	0.729	0.841
21	0.274	0.614	0.763	0.702	0.795	0.664	0.637	0.783	0.882
22	0.615	0.649	0.782	0.313	0.869	0.793	0.470	0.644	0.855
23	0.729	0.648	0.652	0.431	0.715	0.728	0.341	0.707	0.769
24	0.521	0.619	0.639	0.345	0.747	0.740	0.544	0.813	0.886
Mean	0.494	0.692	0.708	0.426	**0.791**	0.738	**0.550**	0.742	**0.818**

Then comparing the contour+normal group and the normal-only group, from the results, the PCK, IoU and N_sim indexes of the contour+normal group are completely better than those of the normal-only group. It can be seen that the contour error can improve the accuracy of the fitting method and can form a collaborative relationship with the surface normal error. In terms of contour alignment, the contour-only group performs best, followed by the contour+normal group. However, in terms of 2D joint alignment, the PCK index of the normal-only group is the worst, with a huge gap of 29.1% with the contour+normal group; meanwhile, the normal similarity index of the contour-only group performs poorly. It can be concluded that although the contour error enables the body shape to be extended to fit the image character region, it lacks local shape alignment and has a negative impact on pose alignment.

3.3.2. The Role of Camera Focus Estimation

This experiment evaluates the effect of focal length estimation on reprojection error, i.e., the effect of weak perspective projection compared to perspective projection. PCK was used as the error metric and two thresholds of 0.05 and 0.1 were used.The experimental data were obtained from different individual images of 3DPW, Human3.6M and MPI-INF-3DHP, the total number of individuals was 14 and 1400 images were selected. The error metric was taken as the average of the corresponding individual image fits and the results are shown in Table 6. It can be seen that all the results obtained using weak perspective projection (with a focal length of 5000 pixels) are poorer than those obtained with perspective projection. At a threshold of 0.05 for PCK, some of the individual fits have a wide range of results. This demonstrates that estimating the camera focal length and using perspective projection correctly under strict alignment conditions can lead to large improvements.

Table 6. Fitting Accuracy using focal length Estimation and using only weak perspective projection

N	Perspective projection		Weak perspective projection	
	PCK 0.05	**PCK 0.1**	**PCK 0.05**	**PCK 0.1**
1	0.472	0.568	0.257	0.558
2	0.381	0.714	0.388	0.746
3	0.366	0.526	0.342	0.589
4	0.394	0.710	0.335	0.550
5	0.388	0.785	0.307	0.613
6	0.442	0.774	0.375	0.668
7	0.365	0.745	0.264	0.581
8	0.363	0.621	0.390	0.666
9	0.410	0.657	0.320	0.703
10	0.395	0.540	0.399	0.560
11	0.483	0.738	0.281	0.573
12	0.316	0.712	0.381	0.652
13	0.402	0.552	0.281	0.639
14	0.347	0.780	0.352	0.508
Mean	**0.395**	**0.673**	0.334	0.615

3.3.3. Role of each module of shape optimization

(1) The role of sub-interval shape optimization

A larger number of different individual images were selected as targets for this experiment, with 6 to 10 images of each sample character. The PCK threshold in the evaluation metric is 0.05. Table 7 demonstrates the ablation experimental data for compartmentalized shape optimization. Individuals numbered 1 to 17 are larger characters, and the others are of standard or thin stature. It can be seen that the fitting results with compartmentalized shape optimization are far superior to those without compartmentalized optimization in terms of joint point alignment. Since the initial value of the shape parameter of is zero, in the absence of interval guidance, the parameters controlling the body shape can only hover around 0 during the optimization process, and the body shape is biased towards the standard body shape. Therefore, as can be seen in the table, the fitting results of the partition-free optimization, the images of the larger body shapes are much less good quality than the interval optimization in both IoU and N_sim. Whereas for standard shaped individuals, IoU and N_sim are similar because the parameter controlling body size is around 0. From the statistical mean data, the mean values of the interval optimization are all smaller than those of the no interval, indicating that the interval optimization strategy performs better in the fitting process.

Table 7. Ablation experimental data on shape optimization between zones

N	Unpartitioned space			Between partitions		
	PCK	**IoU**	**N_sim**	**PCK**	**IoU**	**N_sim**
1	0.407	0.745	0.753	0.453	0.723	0.758
2	0.487	0.763	0.827	0.375	0.731	0.754
3	0.337	0.684	0.724	0.541	0.789	0.835
4	0.261	0.754	0.822	0.408	0.720	0.764
5	0.526	0.797	0.781	0.533	0.842	0.735
6	0.229	0.792	0.728	0.566	0.724	0.834
7	0.321	0.761	0.765	0.428	0.764	0.731
8	0.231	0.729	0.686	0.494	0.851	0.838
9	0.240	0.756	0.788	0.266	0.857	0.863
10	0.537	0.758	0.808	0.300	0.763	0.735
11	0.429	0.770	0.698	0.438	0.766	0.785
12	0.454	0.693	0.718	0.418	0.805	0.727
13	0.252	0.727	0.790	0.548	0.807	0.829
14	0.362	0.763	0.795	0.303	0.840	0.788
15	0.299	0.711	0.813	0.315	0.738	0.811
16	0.527	0.694	0.697	0.481	0.789	0.755
17	0.314	0.722	0.695	0.479	0.823	0.807
18	0.448	0.795	0.706	0.328	0.798	0.785
19	0.460	0.719	0.761	0.347	0.830	0.737
20	0.440	0.743	0.771	0.274	0.805	0.745
21	0.446	0.775	0.725	0.233	0.753	0.836
22	0.370	0.661	0.707	0.568	0.750	0.827
23	0.538	0.670	0.750	0.242	0.747	0.849
24	0.332	0.693	0.718	0.499	0.838	0.838
Mean	0.385	0.736	0.751	**0.410**	**0.785**	**0.790**

(2) Ablation of optimization strategies

In this subsection, the ablation experiment is done again for the optimization strategy module. Figure 5 shows the optimization process for a set of images, where the horizontal axis indicates the number of rounds of iteration, the vertical axis indicates the visualization error, and the curves in the experiments record the best results at the current stage. It can be seen that the final result using only the Adam optimizer has a larger error and also converges the slowest. Adding cosine annealing significantly reduces the error. Combined with subinterval optimization, the error is smaller and converges the fastest, achieving good results after more than 400 iterations.

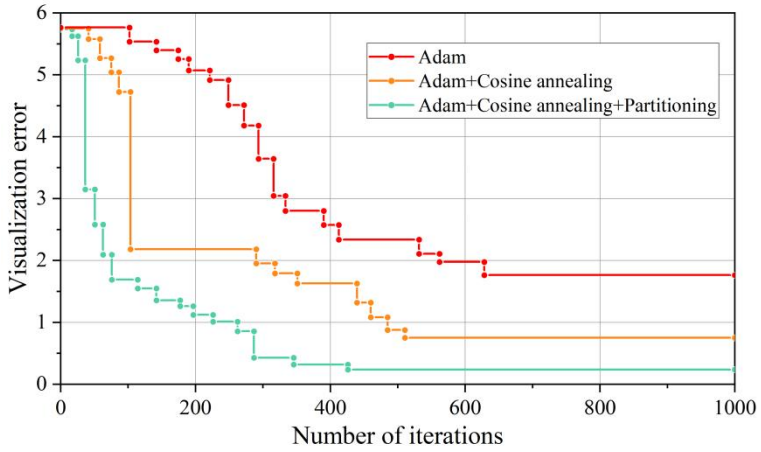

Figure 5. Ablation experiment of shape optimization strategy

4. Discussion

(1) Animation has the value of spreading and passing on national culture

As an important part of cultural soft power, animation bears the responsibility of spreading and inheriting national culture. All along, domestic animation, which is both ideological and cultural, focuses on expressing a correct value orientation through short and concise stories, which is inseparable from China's long-standing creative concepts of “teaching by way of words” and “teaching for fun”. The tradition of strong cultural “didacticism” has its own historical reasons and objective conditions, and the works created under this concept emphasize the value of promoting goodness as beauty, which has played a considerable role in the construction of the moral system of the Chinese nation. The openness and inclusiveness of animation culture can promote the inheritance of traditional culture in a meaningful way.

With the progress of science and technology and the spread of media, the development of animation has moved from a one-way symbolic output stage to a two-way interactive stage, especially the market spread and radiation in many fields such as clothing, food, toys, and so on, which extends the symbolic nature of animation. In the consumer society, animation has received more attention as a cultural product. Animation as a cultural product is firstly embodied in its “commodity”, that is, as a kind of consumption-oriented product, its production and sales are based on the aesthetic choice of consumers, and it constantly adjusts the design of the story, the aesthetics of the picture, the expression of the entertainment in catering to the psychological needs and emotional demands of consumers and with the help of the development of animation derivatives market, it constantly realizes the development of the animation market and the development of the animated products market. With the development of animation derivatives market, the value-added effect of maximizing the consumer market is constantly realized.

Under the tide of globalization and consumer society, Japan and the U.S., by virtue of their global capital operation in economy, science and technology, have been trying their best to expand their position in the cultural industry, and this two-pronged strong attack not only enables them to obtain great economic value, but also brings their values and aesthetics to consumers together in the output of cultural products. The inculcation of these foreign values has created a great threat to the imagination space and value orientation of the local national culture. From the viewpoint of spiritual value culture, the traditional Chinese collective concepts of nation, clan and clans, together with the worship concepts such as loyalty and filial piety, have a unique national form, which is in sharp contrast to the unique spiritual values of the West, such as the concepts of self-sufficiency, libertarianism, and superiority and inferiority, and so on. Therefore, domestic animation can only adhere to the essence of the national spirit in order to spread, presenting a strong cultural soft power.

(2) Cultural identity value of animation

In the context of globalization, different cultures and values intersect and collide with each other. Domestic animation is more likely to produce a sense of identity because its characters show similar or similar thinking patterns, languages, customs and values with Chinese audiences. This sense of identification with animation can prompt Chinese audiences to identify with Chinese culture, create a sense of belonging to Chinese society, and then create cohesion within the social group. For example, in “Magic Brush Ma Liang”, the characters have black eyes and black hair, wear traditional Han Chinese costumes, their houses are straw houses or wooden structures with carved corridors and painted pillars that once existed widely in the Chinese countryside, their agricultural tools are Chinese waterwheels and dragon head sailing boats, the background music has the flavor of Jiangnan silk and bamboo, and the lines and story content also have a strong atmosphere of the times. This series of Chinese elements will undoubtedly bring a strong sense of identity to the audience.

In addition, the naïve and simple Ma Liang is easily recognized by the audience, who not only recognize his values, but also have a sense of belonging to the group he represents. Viewers are also easily bored with the officials and their subordinates who oppress Ma Liang, and then hate the bureaucratic and landlord classes they represent. At the same time, the educational function of animation culture is also obvious. Because, culture and education have never been inseparable, culture has always been an important educational material and tool, and animation culture is no exception.

(3) Artificial intelligence assisted parametric modeling of animation characters

With the development of games, movies, animation and other industries, the demand for animation character modeling has gradually increased. However, modelers manually modeling animation character models requires a lot of time and energy, and the modeling tool itself has a certain threshold for use. In order to solve this problem, this paper uses RaBit parametric model to realize the modeling of 3D animated character models from monocular animated character pictures, which provides a new idea for the existing character modeling process. Specifically, this paper proposes a parametric model regression algorithm based on the two-branch structure and a two-stage cartoon parametric model reconstruction algorithm based on optimization for modeling animated characters. Experiments show that the PSNR values of this paper's method on the three datasets of Render People, THUman 2.0 and Deep Human are 30.83, 29.93 and 31.08, respectively, which are all optimal, and in terms of the rendering efficiency are 64, 48 and 73 FPS, respectively, which achieves a good balance between the reconstruction accuracy and the rendering efficiency.

5. Conclusion

Aiming at the problems in animation character modeling, this paper proposes a parametric model regression algorithm based on the double branch structure, which regresses the morphological parameters and pose parameters of the parametric model from the pictures and models them. In order to solve the difficult problem of parameter regression caused by the varied morphology and exaggerated posture of the animated characters, a two-branch structure is used to regress the morphological parameters and posture parameters respectively. In order to solve the problem of changing morphology of animated characters, the line extraction module is used to guide the algorithm to pay attention to and utilize the line features in the pictures. In order to solve the problem of exaggeration and abstraction of the animated character's pose, geometric information is used as a reference for the predicted pose parameters and fine-tuned several times. In this paper, the parameterized model of animated characters based on the two-branch structure designed by experimental tests on three datasets, namely Render People, THUman 2.0 and Deep Human, significantly improves the reconstruction accuracy and rendering efficiency in comparison with the baseline network, and verifies the feasibility of the model.

References

1. Yusa, I. M. M., Ardhana, I. K., Putra, I. N. D., & Pujaastawa, I. B. G. (2023). Reality in animation: a cultural studies point of view. Eduvest-Journal of Universal Studies, 3(1), 96-109.
2. Tang, M., & Chen, Y. (2024). AI and animated character design: efficiency, creativity, interactivity. The Frontiers of Society, Science and Technology, 6(1), 117-123.
3. Şengünalp, C., & Sarıhan, S. (2024). Convergence of art and technology in character and space design with Blender. Art Time, (7), 38-47.
4. Voci, P. (2023). Para-animation in practice and theory: The animateur, the embodied gesture and enchantment. Animation, 18(1), 23-41.

5. Tan, S. (2022). Animation image art design mode using 3D modeling technology. Wireless Communications and Mobile Computing, 2022(1), 6577461.
6. Guo, Z., Xiang, J., Ma, K., Zhou, W., Li, H., & Zhang, R. (2025). Make-it-animatable: An efficient framework for authoring animation-ready 3d characters. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 10783-10792).
7. Nir, O., Rapoport, G., & Shamir, A. (2022, May). CAST: Character labeling in Animation using Self-supervision by Tracking. In Computer graphics forum (Vol. 41, No. 2, pp. 135-145).
8. Yin, J., & Song, B. (2024). Innovative 3d character model texture mapping solution based on artificial intelligence image generation model. International Journal of Contents, 20(4), 14-21.
9. Luo, Z., Cai, S., Dong, J., Ming, R., Qiu, L., Zhan, X., & Han, X. (2023). Rabit: Parametric modeling of 3d biped cartoon characters with a topological-consistent dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12825-12835).
10. Yu, S., Chen, Y., Guo, T., & Yang, D. (2024). Parametric design algorithm guided by machine vision in animation scene design. Computer-Aided Des. Appl, 21, 261-275.
11. Yang, Z., & Yin, Z. (2021). Efficient hyperparameter optimization for physics-based character animation. Proceedings of the ACM on Computer Graphics and Interactive Techniques, 4(1), 1-19.
12. Eckert, M. L., Um, K., & Thuerey, N. (2019). ScalarFlow: a large-scale volumetric data set of real-world scalar transport flows for computer animation and machine learning. ACM Transactions on Graphics (TOG), 38(6), 1-16.
13. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., & Zhang, L. (2023). Motion-x: A large-scale 3d expressive whole-body human motion dataset. Advances in Neural Information Processing Systems, 36, 25268-25280.
14. Zhang, H., Xu, H., Feng, C., Jampani, V., & Ahuja, N. (2025). Physrig: Differentiable physics-based skinning and rigging framework for realistic articulated object modeling. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6609-6620).
15. Zhang, H., Luo, J., Wan, B., Zhao, Y., Li, Z., Vasilkovsky, M., ... & Zhou, B. (2026). RigMo: Unifying Rig and Motion Learning for Generative Animation. arXiv preprint arXiv:2601.06378.
16. Tan, S., Gong, B., Wang, X., Zhang, S., Zheng, D., Zheng, R., ... & Yang, M. (2024). Animate-x: Universal character image animation with enhanced motion representation. arXiv preprint arXiv:2410.10306.
17. Gui, Z., Xie, J., Han, T., Xie, W., & Zisserman, A. (2025, October). Character-Centric Understanding of Animated Movies. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 3300-3309).
18. Gao, J., Li, J., Liu, W., Zeng, Y., Shen, F., Chen, K., ... & Zhao, C. (2025). CharacterShot: Controllable and Consistent 4D Character Animation. arXiv preprint arXiv:2508.07409.
19. Zhang, N., Meng, H., & Ju, M. (2024). Intelligent construction of animation scenes and dynamic optimization of character images by computer vision. Computer-Aided Design and Applications, 233-246.
20. Bai, Z., Chen, P., Peng, X., Liu, L., Yao, N., & Chen, H. (2024, March). Bring your own character: A holistic solution for automatic facial animation generation of customized characters. In 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR) (pp. 429-438). IEEE.
21. Tang, Y., Guo, J., Liu, P., Wang, Z., Hua, H., Zhong, J. X., ... & Xu, C. (2025). Generative ai for cel-animation: A survey. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3778-3791).
22. Zarif, S., Amin, K. M., Najjar, A., & Wagdy, M. (2025). Animating text descriptions into characters: A comparative review of generative models. IJCI. International Journal of Computers and Information, 12(1), 43-66.

23. Lungu-Stan, V. C., & Mocanu, I. G. (2024). 3D character animation and asset generation using deep learning. Applied Sciences, 14(16), 7234.
24. Qiu, S. (2023). Generative AI processes for 2D platformer game character design and animation. Lecture Notes in Education Psychology and Public Media, 29(1), 146-160.
25. Izani, M., Gabr, M., Razak, A., & Kaleel, A. (2025, July). A Principle-Based Evaluation of Generative AI Animation. In 2025 IEEE Region 10 Symposium (TENSYMP) (pp. 1-7). IEEE.
26. Clocchiatti, A., Fumerò, N., & Soccini, A. M. (2024, January). Character animation pipeline based on latent diffusion and large language models. In 2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR) (pp. 398-405). IEEE.