Scientific Reports

https://doi.org/10.1038/s41598-026-46008-9

Article in Press

Multi-module collaborative 3D human body modeling algorithm based on PIFuHD

Received: 20 February 2026

Accepted: 23 March 2026

Published online: 02 April 2026

Cite this article as: Qiu Z., Zou J., Liu S. *et al.* Multi-module collaborative 3D human body modeling algorithm based on PIFuHD. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-46008-9

Zemin Qiu, Jiajun Zou, Shaojiang Liu & Feng Wang

Multi-module Collaborative 3D Human Body Modeling Algorithm Based on PIFuHD

Zemin Qiu[1], Jiajun Zou[1], Shaojiang Liu [1], Feng Wang[1,*]
[1]School of Information and Intelligence Engineering, Guangzhou Xinhua University, Dongguan, 523133, China
*Corresponding author: Feng Wang (iswf@xhsysu.edu.cn)

Zemin Qiu (qiuzemin@xhsysu.edu.cn)
Jiajun Zou (zoujiajun333@xhsysu.edu.cn)
Shaojiang Liu (mrliuxinhua@xhsysu.edu.cn)
Feng Wang (iswf@xhsysu.edu.cn)

Abstract: To address the limitations of existing 3D human body reconstruction methods in terms of insufficient precision and coarse detail construction, this paper proposes an optimized solution integrating temporal information, human behavior recognition, and multi-module collaboration. The algorithm centers on pixel-aligned hidden functions to establish a full-process reconstruction framework encompassing "temporal constraints-behavioral guidance-feature selection-attitude optimization-3D mapping-detail enhancement". By employing the SAD algorithm for optimal matching point selection and LSTM for temporal dependency capture, the algorithm achieves cross-frame feature coordination. Subsequently, the CNN-LSTM model performs human behavior recognition, using behavioral categories to guide SMPL model's attitude parameter prior and attitude discriminator constraints. Posture normalization eliminates individual variations, while the integration of SMPL model and PIFuHD hidden function enables structured 3D mapping. Finally, octree acceleration grids are utilized to output high-precision 3D human models. Experimental results demonstrate that the proposed algorithm outperforms traditional and literature methods, achieving stable human detail construction in both static standing scenarios and dynamic running scenes.

Keywords: 3D human body modeling; PIFuHD; Human behavior recognition; SMPL

1 Introduction

With the rapid development of virtual reality and human-computer interaction, 3D human body reconstruction technology has become a core method for acquiring spatial information of the human body, with its application scenarios and demands continuously expanding [1-6]. From static image reconstruction to dynamic video sequence reconstruction, researchers are constantly exploring more precise and efficient technical approaches. Existing studies have achieved certain progress in static 3D human pose estimation tasks through architectures such as Hourglass-GCN and dual-stream networks, which integrate skeletal structures and multi-scale features. However, in dynamic scenarios, the temporal correlation of human movements and the diversity of behavioral patterns impose higher requirements on the coherence and consistency of reconstruction results. Current methods still struggle to dynamically balance the contribution of multi-dimensional features to the expression of dynamic details, while their high computational complexity limits real-time deployment [6-12].

In terms of dynamic 3D human body reconstruction, existing algorithms still face the following problems [13-17]:

(1) Most models lack effective modeling of cross frame temporal dependencies, resulting in issues such as pose jumps and motion discontinuities during dynamic sequence reconstruction, which cannot meet the core requirements of reconstruction consistency in dynamic scenes.

(2) The system lacks adaptability to human behavior categories and cannot transform behavioral characteristics into targeted posture constraints. This leads to significant differences in reconstruction accuracy under different motion states.

To address Problem 1, this paper proposes to introduce LSTM networks to capture cross-frame temporal features [18-23], and combines the behavioral recognition results to construct an optimization mechanism, thereby enhancing the modeling of temporal correlations in dynamic sequences.

For the second problem, the behavior categories are transformed into the SMPL model's pose parameter prior and the constraint weights of the pose discriminator. Meanwhile, a multi-dimensional dynamic feature fusion strategy is proposed to dynamically adjust the feature weights based on intra-class consistency and temporal coherence.

Based on the above analysis, this paper proposes a 3D human body reconstruction method that integrates temporal dependence and behavior guidance. The approach incorporates temporal dependence modeling, behavior-guided constraints, and dynamic feature fusion mechanisms into dynamic scene reconstruction tasks, while embedding a closed-loop optimization mechanism of "reconstruction-temporal-behavior". Compared to traditional algorithms, the LSTM network can precisely extract cross-frame temporal correlation features, effectively mitigating pose transition issues. The CNN-LSTM human behavior recognition module learns based on behavioral characteristics, dynamically adjusting pose constraint intensity and discriminator weights to adapt to different behaviors. The multi-dimensional dynamic feature fusion strategy integrates multi-scale, contour-color, and temporal features, enhancing the latent function's ability to express dynamic details.

The contribution of this paper is mainly reflected in the following aspects:

(1) Establish an optimization mechanism that captures cross-frame temporal features via LSTM networks, dynamically adjusts posture constraint intensity based on behavioral recognition results, and significantly improves the coherence of dynamic sequence reconstruction while reducing computational load.

(2) Design a human behavior recognition module to convert behavior categories into SMPL model's pose parameter prior and pose discriminator constraint weights, which improves the accuracy of pose reconstruction and enhances the model's adaptability to different behaviors.

(3) The multi-dimensional feature fusion strategy is proposed to integrate multi-scale features, contour, color features and time series features, and the feature weights are dynamically adjusted based on intra-class consistency and time series coherence to enhance the expression ability of hidden function to dynamic details.

(4) The closed-loop optimization system is established by defining the temporal consistency loss and behavior-guided loss functions, ensuring the continuity of dynamic sequence reconstruction and behavioral consistency.

Based on the above-mentioned time-dependent modeling method and behavior-guided constraint mechanism, this paper constructs a 3D human body reconstruction algorithm that integrates time-dependent and behavior-guided approaches. Experimental results in real-world scenarios demonstrate that the proposed algorithm achieves superior reconstruction accuracy compared to traditional algorithms and those described in the literature.

2 Related Work

2.1 Research on Static 3D Human Body Reconstruction Technology

As the technical cornerstone of dynamic reconstruction, static 3D human body reconstruction aims to accurately restore three-dimensional structures and pose information from single or multiple frames.

Currently, it primarily forms two major technical systems: parametric modeling and non-parametric modeling.

Parametric modeling methods, which describe human body shapes and postures through low-dimensional parameter spaces, have become the mainstream technical framework in the field due to their strong controllability and high data compression ratio. Non-parametric modeling methods, on the other hand, focus on accurately reconstructing complex geometric details of the human body, primarily including techniques such as implicit function modeling, neural radiance field, and 3D Gaussian splashing. The skeletal-aware implicit function model proposed by Pengpeng et al. [27] significantly optimized the spatial mapping logic of implicit functions by introducing skeletal structure prior knowledge, enhancing geometric consistency in single-view human reconstruction. However, its core still focuses on static scenes, failing to consider temporal feature transfer and correlation modeling, making it difficult to directly adapt to dynamic sequence reconstruction needs. Lu et al. explicitly pointed out in their review that recent research in 3D human reconstruction has shifted from early-stage pursuit of static reconstruction accuracy to collaborative optimization of dynamic coherence, complex scene adaptability, and algorithmic efficiency [26]. Yet, the lack of temporal dimension considerations still makes it challenging to meet practical application demands in dynamic scenarios. Many researchers have conducted in-depth studies, but most lack the ability to model temporal dimension information, failing to address the continuity requirements of human motion in dynamic scenes. Moreover, their feature fusion strategies are mostly designed for single-frame data, making it difficult to adapt to the dynamic feature expression needs of multi-frame sequences.

2.2 Temporal-Dependent Modeling for Dynamic 3D Human Body Reconstruction

The core technical challenge in dynamic 3D human body reconstruction lies in accurately capturing cross-frame temporal correlations to ensure the coherence and consistency of motion sequence reconstruction. Current research primarily addresses this issue through two approaches: designing temporal network architectures and optimizing regularization constraints.

In the design of temporal network architectures, recurrent neural networks and their variants have been the mainstream technology for early dynamic modeling. LSTM networks effectively mitigate the gradient vanishing problem through gating mechanisms, enabling precise capture of long-term sequence dependencies. While widely applied in dynamic pose estimation tasks, their serial computation characteristics result in high computational complexity, severely limiting real-time deployment capabilities. In recent years, Transformer architectures have gained attention for their global modeling capabilities through self-attention mechanisms. The HSMR method employs ViT as its backbone network to achieve end-to-end regression from single-frame images to biomechanical skeleton models. However, its temporal modeling relies on frame-interpolation strategies, failing to directly model cross-frame correlation features, which often leads to interpolation artifacts in fast-moving scenarios. To address the prevalent motion blur issue in dynamic scenes, Jing et al. proposed a deep learning-based 3D reconstruction method for motion-blurred images [28], improving reconstruction data quality through image deblurring preprocessing. Nevertheless, this approach did not optimize cross-frame dependency capture mechanisms from the perspective of temporal modeling itself, still exhibiting pose transition issues under rapid motion.

While widely applied in dynamic pose estimation tasks, their serial computation characteristics result in high computational complexity, severely limiting real-time deployment capabilities.

2.3 Research on Behavioral Feature Fusion and Posture Constraints

The diversity of human behavior categories imposes higher requirements on the adaptability of 3D human reconstruction results. Existing research primarily enhances the model's adaptability to different

motion states through two approaches: the fusion of behavior recognition and reconstruction, and the design of pose constraint mechanisms.

In the field of behavior recognition and reconstruction fusion, early approaches predominantly employed a two-stage architecture of "recognize first, reconstruct later". This involved using networks like CNN-LSTM to classify behaviors, then feeding the classification results as fixed weights into the reconstruction model. For instance, Xuhong et al. integrated human pose estimation with behavior analysis techniques for construction worker activity monitoring [24]. By extracting human pose features, they achieved construction behavior classification and safety risk alerts. However, such methods lacked a direct mapping between behavioral features and pose parameters, making it difficult to implement fine-grained pose constraints.

In the design of pose constraint mechanisms, prior constraints of parametric models serve as the core technical approach. Some studies enhance adaptability by optimizing the initialization strategy of SMPL model pose parameters. For instance, Hanif et al. proposed a deep learning fusion framework for gait recognition tasks [25], which improves biometric recognition accuracy by integrating skeletal and appearance features. However, their feature fusion employs a fixed weight allocation mechanism, unable to dynamically adjust the contribution of each feature according to gait complexity. This further highlights the limitations of existing methods in behavioral adaptability.

The existing methods cannot effectively combine behavioral semantics with pose modeling, resulting in insufficient adaptability of 3D human reconstruction results to human behavioral diversity.

2.4 Research on Multi-Dimensional Feature Fusion Strategies

3D human body reconstruction task requires the integration of multi-scale appearance features, structural features and temporal features. The rationality of feature fusion strategy directly affects the reconstruction accuracy and dynamic detail expression ability.

Fixed-weight fusion was the predominant strategy in early research, where feature fusion weights were empirically determined or optimized through grid search. To enhance adaptability, dynamic weight fusion strategies have gained prominence in recent years, with attention mechanisms emerging as the core technology. Tao et al.'s lightweight algorithm [29] improved fusion efficiency through feature channel pruning and adaptive convolution kernel adjustments, yet failed to establish a correlation mechanism between feature weights and behavioral categories or temporal coherence, resulting in insufficient balance between detail restoration and continuity in dynamic scenarios. Additionally, while some studies employ reinforcement learning to dynamically adjust feature weights, the complex training processes and high computational costs severely limit real-time deployment capabilities.

3 Methodology of this paper

3.1 Overall Framework

To address limitations in existing algorithms, this paper proposes a multi-module collaborative reconstruction framework comprising seven core components: Input preprocessing and temporal frame synchronization, optimal matching point selection, human behavior recognition guidance, pose optimization and normalization, multi-dimensional feature fusion, 3D reconstruction and mesh generation, and color mapping. The workflow operates as follows: After preprocessing and synchronization, video frames undergo SAD algorithm screening to identify optimal matching points. The CNN-LSTM model classifies human behavior categories and generates prior pose parameters. Following optimization by the pose discriminator and normalization, the SMPL model performs initial 3D structural modeling. By integrating multi-scale, contour-color, and temporal features, the PIFuHD hidden function predicts SDF values. Finally, octree

acceleration enhances mesh generation and color mapping, producing high-precision, highly coherent 3D human models. The overall algorithm framework is illustrated in Figure 1.

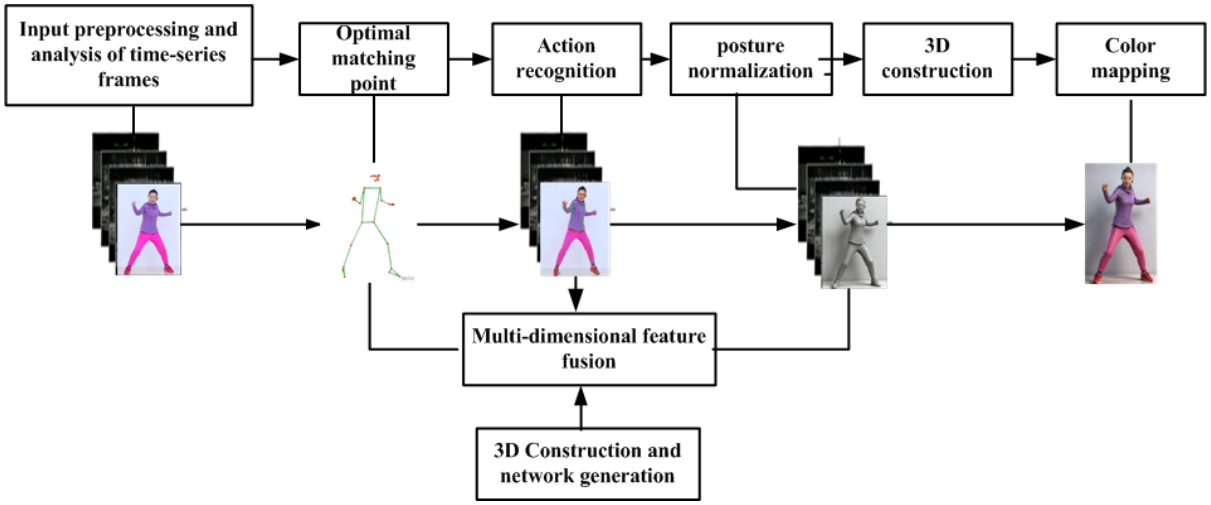


Figure 1 Overall framework of the algorithm

3.2 Input preprocessing and synchronization with time-stamped frames

3.2.1 Image Preprocessing

We perform size normalization and pixel normalization on single-frame images:

$$I_{norm\_size,t} = Resize(I_{in,t},(512,512))$$

$$I_{norm\_pixel,t} = \frac{I_{norm\_size,t}}{255.0} \times 2 - 1 \quad (3)$$

Here, t denotes the frame index, and $I_{in,t}$ represents the original input image of frame t. The system automatically crops the image based on the human detection algorithm.

$$bboxt = HumanDetect(Inorm_pixel,t)$$

$$I_{crop,t} = Crop(I_{norm_pixel,t},bbox_t \times 1.1) \quad (4)$$

3.2.2 Time Sequence Frame Synchronization

For video sequences, we employ timestamp alignment and motion compensation to achieve frame synchronization, ensuring $m_t = (m_{x,t},m_{y,t})$the accuracy of cross-frame feature matching. By defining inter-frame motion vectors, the current frame's cropping position is predicted using the cropped area from the previous frame.

$$bboxt,pred = bboxt\ -1 + m_t \times s \quad (5)$$

Among them, s is the scaling factor, m_t is calculated from the best matching point of adjacent frames, and the overlap rate of the cropped area between frames after synchronization is ≥ 85%.

3.3 Optimal Matching Point Screening and Temporal Feature Extraction

3.3.1 Optimal Matching Point Selection

We employ the SAD algorithm to classify and filter the image blocks in frame t, with the SAD calculation formula as follows:

$$SAD(s_t,c_t(m)) = \sum_{x=1}^{16} \sum_{y=1}^{16} |s_t(x,y) - c_t(x - m_x,y - m_y)| \quad (6)$$

Among them, s_t is the current data of frame t, and c_t is the reference data of frame t. Define the adaptability value SADP, and the optimal matching point determination rule is: SADP_t ≤ 300 □ the optimal matching point in frame t. After filtering, the number of feature points is reduced by 35% to reduce computational redundancy.

After filtering, the number of feature points is reduced by 35%, which reduces the computational redundancy.

3.3.2 Time Series Feature Extraction

This article uses bidirectional LSTM to capture cross frame temporal dependencies. The input is the optimal matching point feature set $F_{match,t-2}, F_{match,t-1}, F_{match,t}$ from frames t-2 to t. The output temporal feature $F_{temp,t}$

$$F_{temp,t} = BiLSTM(F_{match,t-2}, F_{match,t-1}, F_{match,t}) \quad (7)$$

The BiLSTM model features a 256-dimensional hidden layer with an output dimension of 256×64×64, which aligns with the multi-scale feature dimensions.

3.4 Human Behavior Recognition Guidance Module

3.4.1 Behavioral Recognition Model

This article adopts the CNN-LSTM architecture, with the input being the cropped images $I_{crop,t-1}, I_{crop,t}$:
Spatial feature extraction: $F_{spa,t} = ResNet50(I_{crop,t})$, where $F_{spa,t}$, t are single frame spatial features.
Temporal feature fusion: $F_{beh,t} = LSTM(F_{spa,t-1}, F_{spa,t})$, where $F_{beh,t}$ and t are behavioral features.
Behavior classification: $k \in 1,2,...,K$
Among them, k □ 1,2,..., K (K=8, covering common behaviors such as walking, sitting, running, etc.).
Loss function: $L_{cls} = -\sum_{t=1}^{T}\sum_{k=1}^{K} y_{t,k}\log(\hat{y}t,k)$ (8)
Among them, $y_{t,k}$ and k are the true labels of the behavior category in the t-th frame.

3.4.2 Prior Distribution of Attitude Parameters for Behavioral Guidance

We dynamically adjust the SMPL model parameters based on the behavior category (k):
Attitude parameter range: $\Theta_k = [\theta_{k,min}, \theta_{k,max}]$ (9)
Attitude discriminator weights: $\omega_k = \frac{\hat{y}t,k}{\sum k=1^{K}\hat{y}_{t,k}}$, ω_k is the weighted coefficient of behavior confidence, enhancing the focus of current behavioral physiological constraints.
Behavior enhanced posture discriminator
The input of attitude discriminator (D) is "generate virtual joint points+original RGB image+behavioral feature "
The objective function introduces ω_k, and the objective function is $L(G,D) = \omega_k$ □ $[Ex$ □ $Pdata[\log D(x)] + E_{x\square P_G}[\log(1 - D(x))]]$ (10)
Discriminator output:

$$DRII(real,fake,t) = \sigma D(real) - \omega_k \square E_{fake}[D(fake)] \quad (11)$$

$$DRII(fake,real,t) = \sigma D(fake) - \omega_k \square E_{real}[D(real)] \quad (12)$$

The algorithm ensures that the generated attitude conforms to the current human behavior constraints.

Meanwhile, this paper performs displacement and size normalization on 2D joints to eliminate individual differences.

$$S'i,t = Si,t - S_{0,t} \quad (13)$$

$$J_{i,t} = 100 \square \frac{S'i,t}{|S'i,t|} \quad (14)$$

Among them, Si,t are the coordinates of the i-th joint point in the t-th frame, $S_{0,t}$ are the hip center joint points, and the normalized joint point sequence $J_{i,t}$ ranges uniformly from [-1,1].

3.5 Multi-dimensional Feature Fusion Module

3.5.1 Multi-scale Feature Extraction

For the image filtered through optimal matching points, we employ the Hourglass network to extract multi-scale features.

Global features (($F_{g,t} = Hourglass(I_{crop,t}, L = 5)$) .
Local features (($F_{l,t} = Hourglass(I_{crop_256,t}, L = 4)$) , aligned by upsampling to 256 × 64 × 64.

3.5.2 Dynamic Fusion of Triple Features

We integrate multi-scale features, contour-color features, and temporal features, with weights dynamically adjusted based on feature variance and temporal coherence.

$$F_{fusion,t} = \alpha_t \square F_{fusion1,t} + \beta_t \square y_t + \gamma_t \square F_{temp,t} \quad (15)$$

The functions $F_{fusion1,t}$□y_t□γ_t are as follows:

$$F_{fusion1,t} = \frac{Var(F_{g,t})}{Var(F_{g,t}) + Var(F_{l,t})} \square F_{g,t} + \frac{Var(F_{l,t})}{Var(F_{g,t}) + Var(F_{l,t})} \square F_{l,t} \quad (16)$$

$$y_t = (\omega_{1,t} y_{1,t}, \omega_{2,t} y_{2,t})$$

$$\alpha_t + \beta_t + \gamma_t = 1$$

$$\gamma_t = \frac{1}{1 + \exp(-Corr(F_{match,t}, F_{match,t-1}))} \quad (17)$$

$$\omega_{1,t}^2 + \omega_{2,t}^2 = 1$$

We enhance the ability of the hidden function to express spatial details and temporal coherence through triple fusion.

3.6 3D Reconstruction and Mesh Generation

This paper implements a 2D-to-3D structured mapping by integrating the SMPL model, behavioral prior knowledge, and temporal features.

The SMPL model generates a preliminary 3D mesh:

$M_t(\beta,\theta_t) = W(T(\beta,\theta_t), J(\beta,t), \theta_t, W_g)$ Among them, $\theta_t \square \Theta_k$.

3D Joint Coordinate Regression:

$$\boldsymbol{X_{3D,t} = X(M_t(\beta,\theta_t), F_{temp,t})}$$

Implicit Function Extension and SDF Prediction

To extend the pixel alignment implicit function$F_{fusion,t}$, we incorporate multi-dimensional fusion features:

$$f(F_{fusion,t}, X_{3D,t}) = s_t \begin{cases} = 0, & X_{3D,t} \square \text{ surface} \\ 0, & X_{3D,t} \square \text{ external} \\ < 0, & X_{3D,t} \square \text{ internal} \end{cases} \quad (18)$$

We predict the SDF value through the MLP network: $S(p_{i,t}) = MLP(f_{i,t} \square p_{i,t})$.

3.7 Temporal Consistency Loss and Grid Generation

This paper introduces temporal consistency loss to ensure the coherence of adjacent frame models.

$$L_{temp} = \frac{1}{T-1} \sum_{t=2}^{T} \frac{1}{N_v} \sum_{j=1}^{N_v} |v_{j,t} - v_{j,t-1}|2 \quad (19)$$

Among them,$v_{j,t}$ and t are the coordinates of the jth vertex in the t-th frame. And use octree acceleration Marching Cubes algorithm to generate grids.

3.8 Color Mapping Module

This paper implements accurate color sampling through bilinear interpolation algorithm, combining texture inference loss with temporal color consistency loss.

$$L_C = \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} |f(F_{fusion,t}, X_{3D,i,t}) - c(X_{3D,i,t})|^2 + \frac{\lambda}{T-1} \sum_{t=2}^{T} \sum_{i=1}^{n} |c(X_{3D,i,t}) - c(X_{3D,i,t-1})|^2 \quad (20)$$

Among them, $\lambda = 0.1$ is the temporal color weight, ensuring natural color transitions between adjacent frames.

4 Simulation

The proposed algorithm was implemented on the PyCharm platform, with multiple sets of standard videos selected for testing. To validate the algorithm's effectiveness, all experiments were conducted under identical conditions, and the results obtained by this algorithm were compared with those of traditional algorithms and Literature algorithm[27].

The experimental computer configuration includes: Intel Core i9-12900K CPU, 64GB DDR4 RAM, 1TB SSD, and NVIDIA RTX 3090 graphics card with 24GB dedicated memory.

4.1 Comparison of Algorithms

This study quantitatively evaluates the performance of our algorithm and mainstream comparison algorithms in 3D human body reconstruction using three core metrics: point-to-surface distance, Chamfer distance, and 3D IoU. The verified data are presented in Table 1. Point-to-surface distance and Chamfer distance measure geometric errors between reconstructed models and real human models, with smaller values indicating higher reconstruction accuracy. 3DIoU assesses the overlap degree between reconstructed models and real models, where larger values signify better reconstruction consistency and completeness.

Table 1 Experimental Results of Various Algorithms

Configure	Distance from point to surface (mm)	Chamfer distance (mm²)	3D IoU(%)

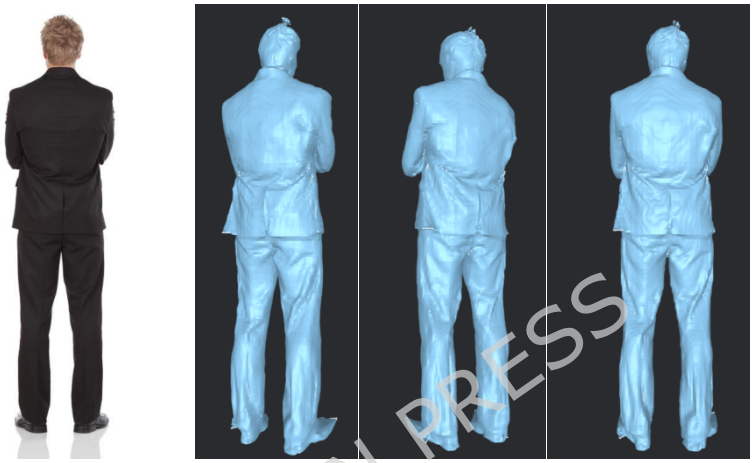
PIFuHD	15.6	0.002	66.0
Literature algorithm[27]	9.9	0.0009	80.8
Algorithm of this paper	9.8	0.0008	81.1

As shown in Table 1, our proposed algorithm achieved significant performance improvements on all three evaluation metrics, outperforming the literature algorithms [27] and the PIFuHD method. This proves the effectiveness of our proposed method.

4.2 Algorithm Experimental Results

To verify the feasibility and effectiveness of the proposed algorithm, two typical motion states were selected for testing: one is the relatively stationary standing state, and the other is the dynamic running process. For these two scenarios, corresponding data were collected respectively, and the 3D detail reconstruction was performed using the algorithm proposed in this paper.

scenario one



(a) Original image (b) 3D model of the character

(c) 3D model (standing)

Figure 2 3D model after color mapping (standing position)

Scene 2:

(a) Original image (b) 3D model of the character

(c) 3D model (running)

Figure 3 3D model after color mapping (running)

The experimental results show that the algorithm can achieve excellent reconstruction performance under various conditions. In addition, Figures 2 and 3 show that human details have been accurately restored through precise color reproduction. In the standing posture scene, the algorithm accurately reconstructs fine clothing details such as subtle wrinkles around the elbows, waist, and knee joints. In the running sequence scene, the reconstructed 3D model maintains significant motion consistency. These effectively demonstrate the advantages of the algorithm in detail construction.

4.3 Ablation Experiment

To validate the functionality of each algorithm module, we conducted ablation experiments and tested them under identical conditions. The test results are presented in Table 2.

Table 2 Ablation Experimental Results

Configure	Distance from point to surface (mm)	Chamfer distance (mm²)	3D IoU□%□
base model	16.6	0.0016	68.0
Basic Model+Optimal Matching Point+Behavior Recognition	12.7	0.0011	74.2
Basic Model + Optimal Matching Point + Temporal Fusion	11.3	0.0012	74.8
Complete method	9.8	0.0009	81.5

As shown in Table 2, when the module selected in this paper is introduced, all indicators improve

compared to the baseline model.

Building upon the "basic model + optimal matching point" framework, the integration of the "behavior recognition" module yielded sustained performance enhancements: the point-to-surface distance was reduced to 12.7mm, the chamfer distance to 0.0011mm², and the 3D IoU improved to 74.2%. This demonstrates that the behavior recognition module significantly improves the matching process, thereby enhancing registration accuracy.

Building upon the "basic model + optimal matching points" framework, the integration of a "temporal fusion" module dramatically reduces the point-to-surface distance to 11.3mm while boosting 3D IoU to 74.8%. This improvement surpasses the "behavior recognition" module in the point-to-surface distance metric, demonstrating that temporal fusion effectively enhances motion accuracy by leveraging inter-frame correlation information.

As shown in Table 2, the complete model achieves optimal performance: the point-to-surface distance is further reduced to 9.8mm, the chamfer distance to 0.0009mm², and the 3D IoU significantly improves to 81.5%. This demonstrates the rationality of each module's design and the effectiveness of their synergistic interaction.

5 Conclusion

We propose a 3D detail reconstruction algorithm integrating temporal information and human behavior recognition, establishing a full-process collaborative framework. The algorithm employs the SAD algorithm to identify optimal matching points and combines LSTM to capture temporal dependencies for cross-frame feature coordination. Based on the CNN-LSTM model, it performs human behavior recognition to guide the SMPL model's pose parameter prior and discriminator constraints. After eliminating individual differences through pose normalization, the SMPL and PIFuHD latent function are fused to achieve structured 3D mapping. Finally, we utilize octree acceleration grids to output high-precision 3D human models. Experimental results demonstrate that the proposed algorithm aligns with expected outcomes, achieving stable 3D detail reconstruction in both standing and running states. Through multi-module collaboration and multi-information fusion, this algorithm provides novel insights for high-precision 3D human reconstruction, though certain limitations remain: the algorithm's robustness in extreme poses or occluded scenes requires further enhancement. Future research will focus on robustness optimization in complex scenarios, lightweight model design, and unsupervised/weakly supervised training methods, aiming to provide comprehensive theoretical support and technical guarantees for practical applications of 3D human reconstruction technology.

Fundings

This research was supported by the 2024 University-level Research Project of Guangzhou Xinhua University (No.2024KYZDZK02), the Guangdong Province Key Construction Discipline Research Capacity Enhancement Project (Nos. 2021ZDJS144,2024ZDJS130), the Characteristic Innovation Category Project of Guangdong Ordinary Colleges and Universities (No. 2024KTSCX127), the School-level Scientific Research Project of Guangzhou Xinhua University (No. 2024KYCXTD02), the Young Innovative Talents Category Project of Guangdong Ordinary Colleges and Universities (Nos. 2023KQNCX124, 2024KQNCX076), and the Key Research Platforms and Projects of Regular Higher Education Institutions in Guangdong Province (No.2025ZDZX3059).

Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author Feng Wang on reasonable request via e-mail iswf@xhsysu.edu.cn.

Competing Interest

The authors declare no competing financial or non-financial interests.

Author Contributions Declaration

Zemin Qiu: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation

Jiajun Zou: methodology, software, validation, formal analysis

Shaojiang Liu: formal analysis, investigation, resources

Feng Wang: writing—review and editing, visualization, supervision, project administration, funding acquisition

Ethics Statements

All datasets and images involved in this study are non-confidential and can be made public. Upon approval by the Institutional Review Board and with reasonable justification, reasonable data access requests can be made to the corresponding author.

Informed Consent: All participants were fully informed of the study purpose, procedures, potential risks, and benefits, and provided written informed consent in accordance with the Declaration of Helsinki. No coercion or inducement was used to obtain consent, and participants retained the right to withdraw from the study at any time without prejudice.

Compliance with Guidelines and Regulations: The study was conducted in accordance with the Declaration of Helsinki, the studies involving human participants were reviewed and approved by School of Information and Intelligence Engineering, Guangzhou Xinhua University Ethics Committee (Approval Number: 21023022023). The participants provided their written informed consent to participate in this study. All methods were performed in accordance with relevant guidelines and regulations.

Author Agreement for Publication

All authors have read and approved the final version of the manuscript, and consent to its publication in the journal.

References

[1] Ange Chen,Chengdong Wu,Chuanjiang Leng.Hourglass-GCN for 3D Human Pose Estimation Using Skeleton Structure and View Correlation[J].Computers, Materials & Continua,2025,82(1):173-191.

[2] Ronghui Wu,Liyun Ma,Zhiyong Chen,Yating Shi,Yifang Shi,Sai Liu,Xiaowei Chen,Aniruddha Patil,Zaifu Lin,Yifan Zhang,Chuan Zhang,Rui Yu,Changyong Wang,Jin Zhou,Shihui Guo,Weidong Yu,Xiang Yang Liu.Stretchable spring-sheathed yarn sensor for 3D dynamic body reconstruction assisted by transfer learning[J].InfoMat,2024,6(4):109-123.

[3] Fenghao Zhang,Lin Zhao,Shengling Li,Wanjuan Su,Liman Liu,Wenbing Tao.3D hand pose and shape estimation from monocular RGB via efficient 2D cues[J].Computational Visual Media,2024,10(1):79-96.

[4] Xiaoxu CAI,Jianwen LOU,Jiajun BU,Junyu DONG,Haishuai WANG,Hui YU.Single depth image 3D face reconstruction via domain adaptive learning[J].Frontiers of Computer Science,2024,18(1):259-261.

[5] Jianchu Lin,Shuang Li,Hong Qin,Hongchang Wang,Ning Cui,Qian Jiang,Haifang Jian,Gongming Wang.Overview of 3D Human Pose Estimation[J].Computer Modeling in Engineering & Sciences,2023(3):1621-1651.

[6] Jee-Sic Hur,Hyeong-Geun Lee,Shinjin Kang,Yeo Chan Yoon,Soo Kyun Kim.3D Face Reconstruction from a Single Image Using a Combined PCA-LPP Method[J].Computers, Materials & Continua,2023(3):6213-6227.

[7] Jian Kang,Wanshu Fan,Yijing Li,Rui Liu,Dongsheng Zhou.3D Human Pose Estimation Using Two-Stream Architecture with Joint Training[J].Computer Modeling in Engineering & Sciences,2023(10):607-629.

[8] Mi Zhou,Rui Liu,Pengfei Yi,Dongsheng Zhou.ER-Net:Efficient Recalibration Network for Multi-View Multi-Person 3D Pose Estimation[J].Computer Modeling in Engineering & Sciences,2023(8):2093-2109.

[9] Cong YU,Dongheng ZHANG,Zhi WU,Zhi LU,Chunyang XIE,Yang HU,Yan CHEN.RFPose-OT:RF-based 3D human pose estimation via optimal transport theory[J].Frontiers of Information Technology & Electronic Engineering,2023,24(10):1445-1457.

[10] LI Chaonan,LIU Sheng,YAO Lu,ZOU Siyu.Video-based body geometric aware network for 3D human pose estimation[J].Optoelectronics Letters,2022,18(5):313-320.

[11] Chao Yang,Xuyu Wang,Shiwen Mao.RFID-based 3D human pose tracking: A subject generalization approach[J].Digital Communications and Networks,2022,8(3):278-288.

[12] Xiaoxing Zeng,Zhelun Wu,Xiaojiang Peng,Yu Qiao.Joint 3D facial shape reconstruction and texture completion from a single image[J].Computational Visual Media,2022,8(2):239-256.

[13] Yanlong Tang,Yun Zhang,Xiaoguang Han,Fang-Lue Zhang,Yu-Kun Lai,Ruofeng Tong.3D corrective nose reconstruction from a single image[J].Computational Visual Media,2022,8(2):225-237.
[14] Peng Jin,Shaoli Liu,Jianhua Liu,Hao Huang,Linlin Yang,Michael Weinmann,Reinhard Klein.Weakly-Supervised Single-view Dense 3D Point Cloud Reconstruction via Differentiable Renderer[J].Chinese Journal of Mechanical Engineering,2021,34(5):195-205.
[15] Khalil Khan,Jehad Ali,Kashif Ahmad,Asma Gul,Ghulam Sarwar,Sahib Khan,Qui Thanh Hoai Ta,Tae-Sun Chung,Muhammad Attique.3D Head Pose Estimation through Facial Features and Deep Convolutional Neural Networks[J].Computers, Materials & Continua,2021(2):1757-1770.
[16] Xianhua Li,Haohao Yu,Shuoyu Tian,Fengtao Lin,Usama Masood.Multi-Branch High-Dimensional Guided Transformer-Based 3D Human Posture Estimation[J].Computers, Materials & Continua,2024,78(3):3551-3564.
[17] Qiqi He,Li Li,Dai Li,Tao Peng,Xiangying Zhang,Yincheng Cai,Xujun Zhang,Renzhong Tang.From Digital Human Modeling to Human Digital Twin: Framework and Perspectives in Human Factors[J].Chinese Journal of Mechanical Engineering,2024,37(1):1-14.
[18] Ameni Ellouze,Nesrine Kadri,Alaa Alaerjan,Mohamed Ksantini.Combined CNN-LSTM Deep Learning Algorithms for Recognizing Human Physical Activities in Large and Distributed Manners:A Recommendation System[J].Computers, Materials & Continua,2024,79(4):351-372.
[19] Navaneetha Krishnan Muthunambu,Senthil Prabakaran,Balasubramanian Prabhu Kavin,Kishore Senthil Siruvangur,Kavitha Chinnadurai,Jehad Ali.A Novel Eccentric Intrusion Detection Model Based on Recurrent Neural Networks with Leveraging LSTM[J].Computers, Materials & Continua,2024,78(3):3089-3127.
[20] Yi-Chun Lai,Shu-Yin Chiang,Yao-Chiang Kan,Hsueh-Chun Lin.Coupling Analysis of Multiple Machine Learning Models for Human Activity Recognition[J].Computers, Materials & Continua,2024,79(6):3783-3803.
[21] Jiajie Shen,Yan Wang,Dongxu Zhang.A Novel Locomotion Rule Embedding Long Short-Term Memory Network with Attention for Human Locomotor Intent Classification Using Multi-Sensors Signals[J].Computers, Materials & Continua,2024,79(6):4349-4370.
[22] Meng Zhu,Xiaorong Guan,Zhong Li,Long He,Zheng Wang,Keshu Cai.sEMG-Based Lower Limb Motion Prediction Using CNN-LSTM with Improved PCA Optimization Algorithm[J].Journal of Bionic Engineering,2023,20(2):612-627.
[23] Alaa Omran Almagrabi.A Deep CNN-LSTM-Based Feature Extraction for Cyber-Physical System Monitoring[J].Computers, Materials & Continua,2023,76(8):2079-2093.
[24] Xuhong Zhou,Shuai Li,Jiepeng Liu,Zhou Wu,Yohchia Frank Chen.Construction Activity Analysis of Workers Based on Human Posture Estimation Information[J].Engineering,2024,33(2):225-236.
[25] Ch Avais Hanif,Muhammad Ali Mughal,Muhammad Attique Khan,Nouf Abdullah Almujally,Taerang Kim,Jae-Hyuk Cha.Human Gait Recognition for Biometrics Application Based on Deep Learning Fusion Assisted Framework[J].Computers, Materials & Continua,2024,78(1):357-374.
[26] Lu Chen,Sida Peng,Xiaowei Zhou.Towards efficient and photorealistic 3D human reconstruction: A brief survey[J].Visual Informatics,2021,5(4):11-19.
[27] Pengpeng Liu,Guixuan Zhang,Shuwu Zhang,Yuanhao Li,Zhi Zeng.Skeleton-aware implicit function for single-view human reconstruction[J]. CAAI Transactions on Intelligence Technology, 2023,8(2):379-389.
[28] Jing Zhang,Keping Yu,Zheng Wen,Xin Qi,Anup Kumar Paul.3D Reconstruction for Motion Blurred Images Using Deep Learning-Based Intelligent Systems[J].Computers, Materials & Continua,2021(2):2087-2104.
[29] Tao Zhang,Yi Cao.Improved Lightweight Deep Learning Algorithm in 3D Reconstruction[J]. Computers, Materials & Continua,2022(9):5315-5325.