IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1479-1494
Journal homepage: http://iieta.org/journals/ts

A Multi-View Image Registration-Integrated Method for Three-Dimensional Human Motion Reconstruction in Physical Education

Xue Deng, Xue Bai*, Zepei Li

Sports Work Department, Hebei Vocational University of Technology and Engineering, Xingtai 054000, China

Corresponding Author Email: baixue@hevute.edu.cn

https://doi.org/10.18280/ts.430331

Received: 20 October 2024
Revised: 13 December 2024
Accepted: 11 February 2025
Available online: 30 June 2026

Keywords:
multi-view image registration, three-dimensional human reconstruction, human pose estimation, skinned multi-person linear model, quantitative motion assessment, intelligent physical education

ABSTRACT

Three-dimensional human motion reconstruction constitutes a core technical approach for quantitative analysis and movement correction in intelligent physical education instruction. As for existing multi-view reconstruction methods, reconstruction accuracy and temporal stability remain difficult to guarantee under scenarios characterized by high-speed limb movements and frequent self-occlusions. To address the above issues, a multi-view image registration-integrated method for three-dimensional human motion reconstruction in physical education was proposed, and an end-to-end technical framework unifying registration, reconstruction, and evaluation was constructed. By leveraging human joint semantic priors and incorporating graph attention networks, cross-view collaborative feature matching was realized, effectively mitigating cross-view matching ambiguities under complex motion and occlusion conditions. Temporal motion smoothness constraints and an extended Kalman filter-based dynamic view selection mechanism were introduced to suppress inter-frame jitter in the reconstructed three-dimensional sequences, while simultaneously optimizing inference efficiency without compromising reconstruction precision. A multi-dimensional quality assessment system, encompassing pose confidence, occlusion degree, and viewing angles, was established to enable adaptive multi-view feature fusion and optimization of the skinned multi-person linear model parameters. On this basis, a quantitative evaluation module tailored for physical education teaching was developed, enabling visual feedback of human motion deviations. Systematic comparative experiments and ablation validations were conducted on both public benchmark datasets and a self-constructed real-world physical education motion dataset. Results demonstrated that the proposed method significantly outperformed existing state-of-the-art algorithms in terms of image registration accuracy and three-dimensional pose reconstruction performance, while effectively improving the temporal smoothness of motion sequences. Pedagogical user testing confirmed that the method was capable of precisely quantifying motion deficiencies and intuitively presenting motion deviations, rendering it highly suitable for physical education classroom instruction and specialized training analysis. This study is expected to provide efficient and reliable technical support for intelligent physical education instruction, quantitative human motion assessment, and sports science analytics.

1. INTRODUCTION

Digital and intelligent technologies are profoundly empowering the transformation and upgrading of traditional physical education. Precise quantitative analysis of human motion constitutes the foundational basis for scientific pedagogical guidance, standardized movement error correction, and sport-specific performance assessment [1, 2]. Three-dimensional human motion reconstruction technology is capable of transcending the planar observational limitations of two-dimensional imagery, enabling accurate restoration of spatial postures, motion trajectories, and deformation patterns of human limb movements. It serves as the core enabling technology bridging computer vision and intelligent physical education instruction, with widespread applications in sports science analytics, rehabilitation training correction, and human-computer interaction [3, 4]. In realistic physical education scenarios, typical sports actions such as basketball shooting, standing long jumps, and gymnastic flips are characterized by high motion velocity, substantial limb amplitude variations, and frequent self-occlusions between the torso and extremities. Traditional single-view three-dimensional reconstruction methods are inherently limited in their ability to resolve depth ambiguities, observational blind zones, and occlusion interferences, rendering the stability and accuracy of reconstruction outcomes insufficient for pedagogical analytical standards [5, 6]. Multi-view synchronous imaging systems can effectively mitigate depth ambiguity problems inherent in single-view imaging through complementary multi-dimensional scene geometric information, and cross-view feature cross-validation can be leveraged to achieve pose inference in occluded regions [7, 8]. Concurrently, the rapid advancement of parametric human models has driven the evolution of human reconstruction from

sparse joint localization to full-body mesh modeling, substantially enhancing the completeness and realism of three-dimensional human representations [9, 10]. Nevertheless, existing technical frameworks are predominantly designed for general static scenarios and have failed to accommodate the specific requirements of physical education contexts, which demand not only high-precision three-dimensional pose recovery but also quantifiable and visualizable motion evaluation capabilities, resulting in notable deficiencies in technical deployability and pedagogical practicality [11, 12].

Significant technical deficiencies persist in current multi-view three-dimensional human motion reconstruction frameworks, rendering them inadequate for meeting the reconstruction and pedagogical evaluation requirements of dynamic sports motions. In prevailing mainstream algorithms, cross-view image registration and three-dimensional human model fitting are commonly decomposed into two independent serial execution modules, thereby severing the geometric coupling between feature matching and three-dimensional reconstruction. Feature matching errors introduced during the registration phase cannot be compensated or corrected through the model optimization process, nor can the cross-view geometric constraint information extracted during registration be adequately utilized to serve three-dimensional pose optimization, resulting in an error accumulation effect that constrains reconstruction accuracy for complex sports motions [13, 14]. At the level of cross-view information fusion, existing methods lack dynamic perceptual capabilities regarding view-specific imaging quality. Most approaches accomplish multi-view feature fusion through uniform weighting or fixed strategies, without differentiating the informational contribution disparities among distinct viewpoints [15, 16]. The dynamic variations inherent in sports motions induce real-time fluctuations in imaging quality across different views, with certain viewpoints being prone to extensive occlusion and loss of effective features. Fixed fusion paradigms introduce invalid noise features, thereby diminishing reconstruction robustness under dynamic scenarios [17, 18]. In temporal dimension modeling, current studies predominantly adopt frame-independent reconstruction processing, neglecting the inherently continuous evolutionary nature of sports motions. Inter-frame motion correlations and trajectory continuity constraints are not leveraged to optimize reconstruction outcomes, leading to severe inter-frame jitter and pose discontinuities in the output motion sequences, which significantly compromises the efficacy of motion playback observation and dynamic pedagogical analysis [19, 20]. More critically, existing investigations predominantly concentrate on optimizing the reconstruction accuracy of three-dimensional human models, with the output of three-dimensional skeletons and mesh models being treated as the sole research objective. Standardized pedagogical evaluation criteria have not been integrated to construct quantitative analysis frameworks, and standardized instructional metrics cannot be extracted from reconstruction results, resulting in a pronounced disconnect between technical research and practical teaching applications [21, 22].

To address the aforementioned core limitations and application pain points of existing technologies, a unified three-dimensional human motion reconstruction framework integrating registration, reconstruction, and evaluation is constructed, with full-process technical optimization and adaptation being carried out for physical education instructional scenarios. Human semantic priors and graph neural networks are combined to achieve cross-view collaborative feature matching, thereby resolving registration ambiguities under dynamic occlusion and viewpoint variations. Temporal smoothness constraints and a dynamic view selection mechanism are introduced to balance reconstruction temporal stability and algorithmic inference efficiency. A multi-dimensional view confidence evaluation system is established to enable adaptive multi-view feature fusion and human model parameter optimization. A quantitative motion evaluation and visual feedback module is developed to realize deep integration of three-dimensional reconstruction technology with intelligent physical education instruction. Experimental results on multiple benchmark datasets and a self-constructed physical education motion dataset fully validate the technical superiority and pedagogical practicality of the proposed approach.

The overall organizational structure is clear and progressively developed. In Chapter 2, the overall framework and core technical principles of the proposed method are systematically elaborated, with the design rationale and implementation workflows of the four functional modules being described in detail. The technical innovations and optimization logic of each module are clarified in conjunction with existing research deficiencies. In Chapter 3, the effectiveness and advancement of the proposed method are validated through ablation experiments, comparative experiments, and user surveys from multiple dimensions, including quantitative metrics, qualitative effects, and practical applications. In Chapter 4, the research outcomes are comprehensively summarized, the limitations of the current study are delineated, and future directions for optimization and application scenarios of intelligent sports reconstruction technology are prospected.

2. METHODOLOGY

To systematically address the critical issues inherent in existing multi-view three-dimensional human reconstruction methods—including poor modular coupling, inadequate view-adaptive capacity, insufficient utilization of temporal motion features, and weak compatibility with pedagogical applications—an integrated reconstruction and quantitative evaluation technical framework tailored for dynamic sports motion scenarios is constructed. A closed-loop, full-process pipeline from multi-view image inputs to pedagogical motion feedback is realized through four progressively structured functional modules. The entire approach takes high-precision image registration as its core entry point. First, cross-view geometric association information is mined through a human semantic-driven collaborative feature matching strategy, from which highly reliable cross-view matching correspondences are output, thereby effectively overcoming registration deviations caused by high-speed sports motions and limb occlusions. On this basis, a temporal motion constraint mechanism is introduced to accomplish feature temporal propagation and dynamic view screening, effectively suppressing pose jitter induced by frame-independent reconstruction while balancing reconstruction accuracy and computational efficiency. The method further incorporates multi-dimensional view quality assessment metrics to achieve adaptive weighted feature fusion, and a more robust three-dimensional human motion sequence is output through an

iterative optimization strategy of the parametric human model. Based on the high-precision motion data obtained from reconstruction, the method ultimately accomplishes extraction of sports motion keyframes, quantification of multi-dimensional motion parameters, and visual analysis of deviations from standard motions, thereby realizing deep integration of three-dimensional vision reconstruction technology with physical education instructional scenarios. In this chapter, the design rationale, technical principles, and specific implementation approaches of each core module are elaborated layer by layer, providing a comprehensive presentation of the overall technical framework of the proposed method.

2.1 Human semantic-aware multi-view collaborative feature matching

Conventional cross-view feature matching methods predominantly accomplish feature association based on two-view independent processing, which renders them inadequate for complex scenarios involving dynamic deformations and frequent self-occlusions characteristic of sports motions. Local visual information from a single viewpoint is highly prone to inducing matching ambiguities and erroneous correspondences. To enhance registration robustness under dynamic human body scenarios, a dual-branch feature extraction architecture is constructed, in which human pose semantic information and deep visual features are simultaneously mined. Feature encoding is accomplished through dual backbone networks respectively, with human joint heatmaps being output by a pre-trained visual Transformer network to characterize the spatial probability distributions of human key body parts, while scale-robust visual features are extracted using a high-resolution network. Spatial max-pooling is applied to the full-channel joint heatmaps, from which semantic attention maps focusing on human body regions are generated. Semantic-adaptive weighted enhancement of visual features is realized through normalization operations, with the specific computation expressed as:

$$\widetilde{F}_i = F_i \odot softmax(A_i/\tau_a) \tag{1}$$

where, F_i denotes the original visual features, A_i represents the semantic attention map, τ_a is set to 0.5 to regulate the focusing intensity of the attention distribution, and $\odot$ denotes element-wise feature weighting. This encoding mechanism automatically enhances feature weights in joint-critical regions while suppressing invalid interferences from complex backgrounds and clothing textures, thereby providing stable semantic prior support for accurate cross-view matching. Figure 1 illustrates the flowchart of the human semantic-aware multi-view collaborative feature matching process.

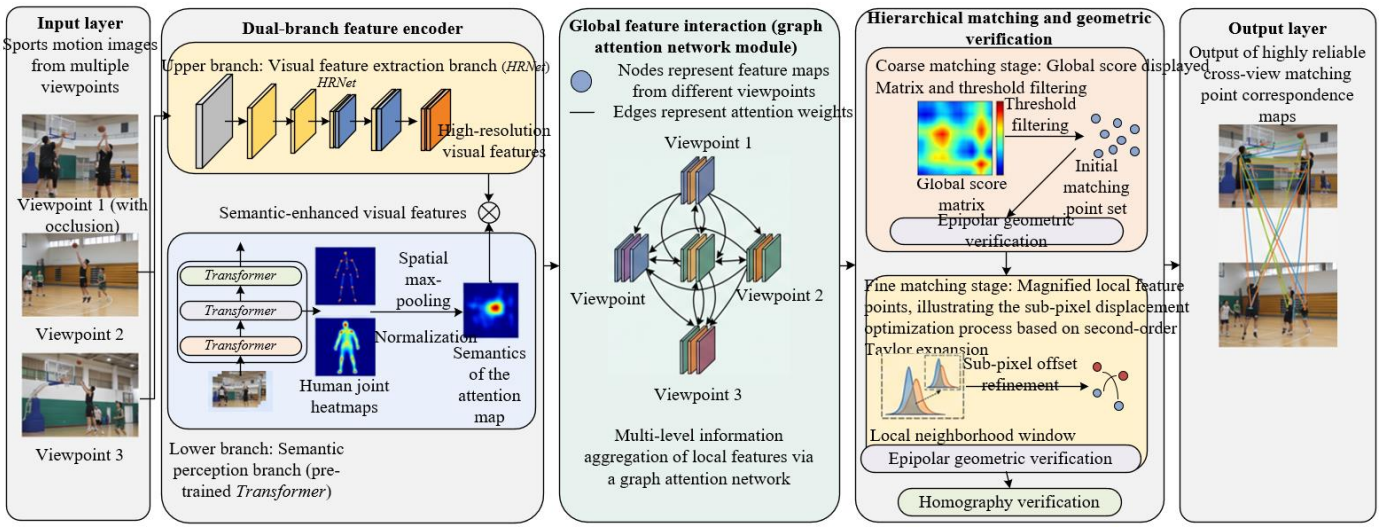


Figure 1. Flowchart of human semantic-aware multi-view collaborative feature matching

To overcome the informational limitations inherent in two-view matching, multiple camera viewpoints are organized into a complete graph topological structure, with feature maps from each viewpoint serving as network nodes, and a cross-view global feature association mechanism is established. Based on the principle of differentiable feature matching, original features are mapped into a unified metric space through learnable projection matrices, and an association score matrix between pixels across different viewpoints is computed as:

$$S_{ij} = softmax_{row}\left(\frac{(W_Q\widetilde{F}_i)(W_K\widetilde{F}_j)^{\top}}{\sqrt{d}}\right) \tag{2}$$

where, W_Q and W_K denote the query and key projection matrices, respectively, and d represents the unified feature dimension. The score matrix quantifies the matching probabilities between pixels across different viewpoints. On this basis, a dynamic graph attention network is introduced to accomplish multi-level global information interaction, through which complementary feature information from all adjacent viewpoints is aggregated via a multi-head attention mechanism. The iterative feature update process is expressed as:

$$\widetilde{F}_i^{(l+1)} = \widetilde{F}_i^{(l)} + FFN\left(AGG_{k=1}^{K_{head}} \sum_{j\in N(i)} \alpha_{ij}^{k,(l)} W_V^{k,(l)} \widetilde{F}_j^{(l)}\right) \tag{3}$$

The attention coefficient $\alpha_{ij}^{k,(l)}$ is adaptively solved through a dynamic activation function and feature concatenation operations, by which information fusion weights are dynamically adjusted according to the feature association quality across different viewpoints. After four layers of

iterative graph network inference, single-view features are sufficiently fused with global multi-view geometric and semantic information, effectively mitigating matching uncertainties induced by dynamic occlusions and viewpoint distortions.

To balance matching efficiency with sub-pixel localization accuracy, a coarse-to-fine hierarchical matching strategy is adopted for feature point pair screening and refinement. In the coarse matching stage, spatial local maxima are extracted based on the global association score matrix, and an initial matching point set is filtered with a fixed confidence threshold, through which low-relevance invalid matching combinations are rapidly eliminated, substantially reducing the computational overhead of subsequent fine optimization. In the fine refinement stage, for the feature point pairs obtained from coarse matching, a second-order Taylor expansion model is constructed within local neighborhood windows, and the pixel offset correction is derived by solving the second-order partial derivative matrix of the feature distribution:

$$\delta p=-\left(\frac{\partial^2 S_{ij}}{\partial p^2}\right)^{-1}\frac{\partial S_{ij}}{\partial p} \tag{4}$$

Matching coordinates are iteratively updated until parameter convergence, thereby achieving sub-pixel accuracy optimization of matching positions. This hierarchical optimization scheme avoids redundant computations associated with global fine matching while effectively correcting pixel offset errors induced by high-speed sports motions, significantly enhancing localization accuracy in cross-view feature matching.

To ensure that the matching results comply with the geometric constraint principles of the multi-camera system, a dual-verification mechanism incorporating epipolar geometry and homography is introduced for the final screening of the matching point set. Based on the camera calibration parameters, the fundamental matrices between different viewpoints are solved, and a bidirectional symmetric epipolar distance constraint criterion is constructed as:

$$d_{epip}(p_i,p_j)=\frac{|p_j^{\mathrm{T}}F_{ij}p_i|}{\sqrt{(F_{ij}p_i)_1^2+(F_{ij}p_i)_2^2}}+\frac{|p_i^{\mathrm{T}}F_{ji}p_j|}{\sqrt{(F_{ji}p_j)_1^2+(F_{ji}p_j)_2^2}}<\tau_{epip} \tag{5}$$

where, F_{ij} denotes the fundamental matrix between viewpoints, and τ_{epip} is set to 2.5 pixels to constrain the epipolar projection error of matching points. A homography matrix reprojection verification rule is simultaneously introduced to further eliminate outlier point pairs with excessively large geometric mapping deviations. Following multi-level screening and optimization, the module ultimately outputs highly reliable cross-view matching point sets, corresponding matching confidences, and relative pose parameters between camera viewpoints, thereby providing an accurate geometric matching foundation for subsequent temporal constraint optimization, multi-view feature fusion, and human model parameter solving.

2.2 Motion-aware temporal consistency constraints and dynamic registration

To fully exploit the inherent temporal continuity characteristics of sports motions and to address the issues of matching redundancy and dynamic ambiguities inherent in frame-independent registration, an inter-frame dense feature propagation mechanism is constructed, through which adaptive registration initialization driven by temporal priors is realized. Figure 2 illustrates the principles of the temporal consistency constraints and dynamic view-adaptive registration mechanism. Based on the motion correlation properties of consecutive video frames, pixel-level temporal motion offset patterns are modeled. The feature matching search range for subsequent frames is constrained using forward optical flow fields, through which the high computational cost of global exhaustive matching is effectively avoided. The temporal adaptive search region update rule is formulated as:

$$R_{t+1}(p)=R_t(p+\Phi_{t\to t+1}(p))+N(0,\sigma^2 I) \tag{6}$$

where, $\Phi_{t\to t+1}$ denotes the dense optical flow field between adjacent frames, which is used to precisely characterize pixel-level temporal motion trajectories, and N represents a two-dimensional Gaussian perturbation term with a standard deviation of 3 pixels, which is introduced to cover motion prediction deviations and ensure completeness of the search region. For scenarios where optical flow estimation fails—such as motion boundaries and limb occlusions—the algorithm automatically reverts to a global search strategy, thereby guaranteeing matching effectiveness under extreme dynamic conditions. A forward-backward optical flow consistency verification mechanism is simultaneously introduced, with an error threshold of 1.5 pixels being set to filter highly reliable optical flow features, through which temporally anomalous matching results are further eliminated. This mechanism reduces matching computational complexity from quadratic to linear order, substantially improving dynamic registration efficiency, while cross-view matching ambiguities in high-speed sports motions are effectively mitigated through the temporal motion constraints.

To fundamentally resolve the issues of pose discontinuities and sequence jitter caused by frame-independent reconstruction, a second-order temporal smoothing regularization constraint is introduced, through which a trajectory optimization mechanism conforming to human motion patterns is constructed. Based on the three-dimensional joint coordinates of consecutive temporal frames, a motion acceleration field is constructed to quantify the instantaneous kinematic variation characteristics of human limbs, by which the dynamic evolutionary patterns of sports motions are precisely characterized. The acceleration field is computed as:

$$a_k^{(t)}=X_k^{(t-1)}-2X_k^{(t)}+X_k^{(t+1)} \tag{7}$$

Based on the acceleration field, a weighted least-squares smoothing loss function is constructed to achieve temporal constraint optimization over the entire motion sequence, with the specific expression given as:

$$L_{smooth}=\sum_{t=2}^{T-1}\sum_{k=1}^{K}\gamma_k\cdot\left\|X_k^{(t-1)}-2X_k^{(t)}+X_k^{(t+1)}\right\|_2^2 \tag{8}$$

where, $X^{(t)}_k$ denotes the three-dimensional coordinates of the k-th human joint at frame t, and γ_k represents the joint-adaptive weight coefficient. In accordance with the physiological characteristics of human motion, higher weights are assigned to distal joints—such as wrists and ankles—which possess

greater degrees of motion freedom, while lower weights are assigned to core joints—such as the torso and pelvis—which exhibit greater stability, thereby accommodating the motion differences across different body parts. The solution matrix of this regularization term exhibits a pentadiagonal banded structure, which enables efficient iterative solving through Cholesky decomposition, with the overall computational complexity being linear with respect to the total number of video frames, thereby maintaining high algorithmic efficiency while achieving high-precision temporal smoothing.

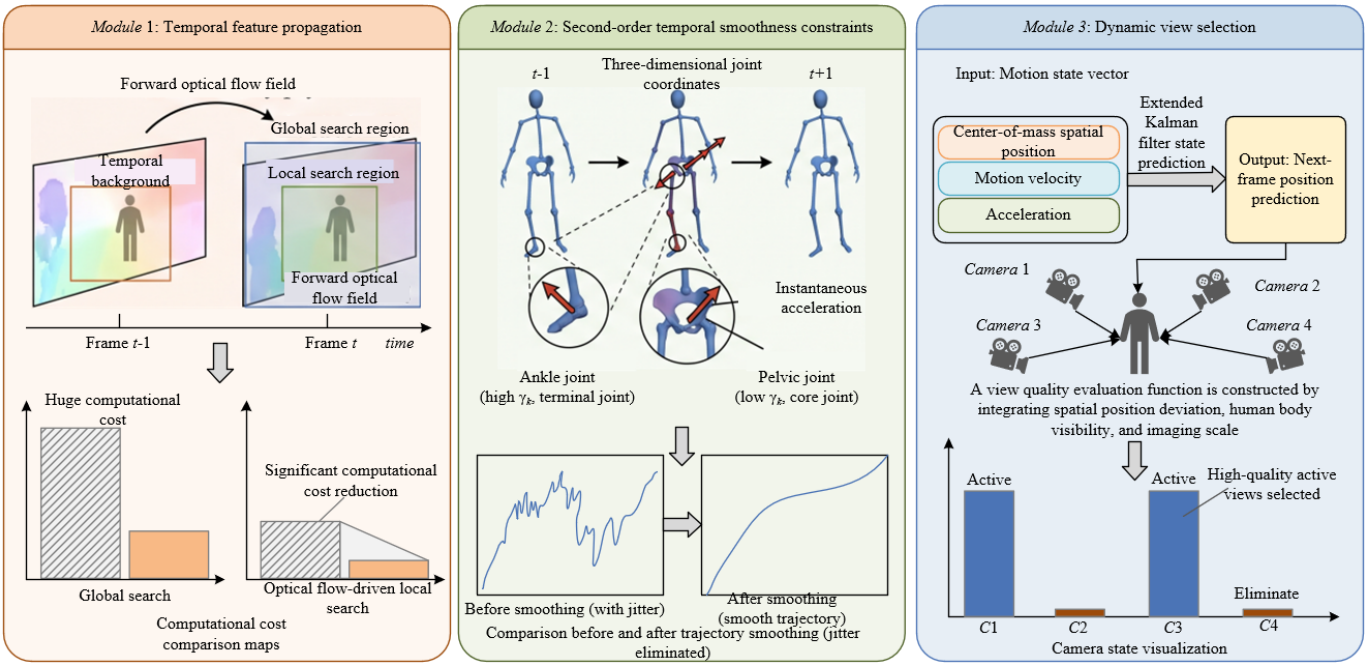


Figure 2. Temporal consistency constraints and dynamic view-adaptive registration mechanism

To accommodate the dynamic characteristics of rapid movement and instantaneous posture changes in sports motions, while balancing three-dimensional reconstruction accuracy and inference efficiency, real-time prediction of human motion states is realized through an extended Kalman filter, and dynamic view-adaptive screening is accomplished through the establishment of multi-dimensional quantitative criteria. A nine-dimensional motion state vector is constructed by integrating the spatial position of the human center of mass, motion velocity, and acceleration, and accurate prediction of human motion trends is achieved through temporal state transition equations. The core prediction formula is given as:

$$\hat{s}^{(t|t-1)}=A\cdot s^{(t-1)}+w \tag{9}$$

where, A denotes the temporal state transition matrix, with the temporal update relationship being constructed according to the camera sampling frequency of 30 fps, and w represents the Gaussian process noise, which is used to fit the uncertainty associated with non-uniform human motion. The predicted three-dimensional human center of mass is projected onto each view's image plane, and a view quality evaluation function is constructed by integrating spatial position deviation, human body visibility, and imaging scale:

$$q_i^{(t)}=exp\left(-\frac{\|\hat{x}_i^{(t)}-c_i\|_2^2}{2\sigma_c^2}\right)\cdot I(vis_i^{(t)}>\tau_{vis})\cdot\left(1+\frac{scale_i^{(t)}}{\max\limits_i scale_i^{(t)}}\right) \tag{10}$$

where, c_i denotes the image center coordinate of the view, σ_c regulates the spatial position penalty weight, $vis_i^{(t)}$ characterizes the effective visible proportion of the human body, and $scale_i^{(t)}$ characterizes the projected imaging area of the human body. Through this evaluation function, the real-time informational contribution capacity of each viewpoint is quantified, high-quality active views are selected for subsequent reconstruction, and low-quality views with severe occlusion and insufficient effective information are discarded, thereby enabling adaptive allocation of view resources under dynamic scenarios and achieving balanced optimization between reconstruction accuracy and computational efficiency.

2.3 Confidence-guided adaptive multi-view feature fusion and skinned multi-person linear parameter optimization

To address the issue of imbalanced information quality across different viewpoints in multi-view imaging, a multi-dimensional quantification system is established for precise assessment of view-specific imaging quality, thereby providing adaptive weighting foundations for subsequent feature fusion and model optimization. Figure 3 illustrates the workflow of confidence-guided three-dimensional feature fusion and skinned multi-person linear model optimization. Three complementary evaluation dimensions—pose detection confidence, limb occlusion degree, and spatial observation angle—are integrated, and a view comprehensive confidence computation model is constructed to achieve dynamic adaptive solving of each view's contribution at every frame. The overall computation is expressed as:

$$c_i^{(t)}=\alpha\cdot c_i^{pose(t)}+\beta\cdot c_i^{occ(t)}+\gamma\cdot c_i^{angle(t)} \tag{11}$$

where, the weight parameters are determined through grid search on the validation set, with α, β, and γ set to 0.35, 0.35, and 0.30, respectively. The pose confidence is computed through weighted aggregation of multi-joint detection scores,

with weights assigned according to limb motion amplitudes, thereby enhancing the evaluation proportion of dynamically moving joints. The occlusion confidence is derived from the proportion of effectively visible human body regions obtained from semantic segmentation results, with the output weights of heavily occluded views being suppressed through an exponential decay function. The viewing angle confidence is adaptively assigned based on the spatial angle between the human body orientation and the camera viewing direction, ensuring that perpendicular viewing angles are granted higher weight advantages. Following the fusion of multi-dimensional scores, global view confidences are subjected to normalization, while invalid information inputs from inactive views are simultaneously masked, thereby achieving dynamic and precise allocation of view weights.

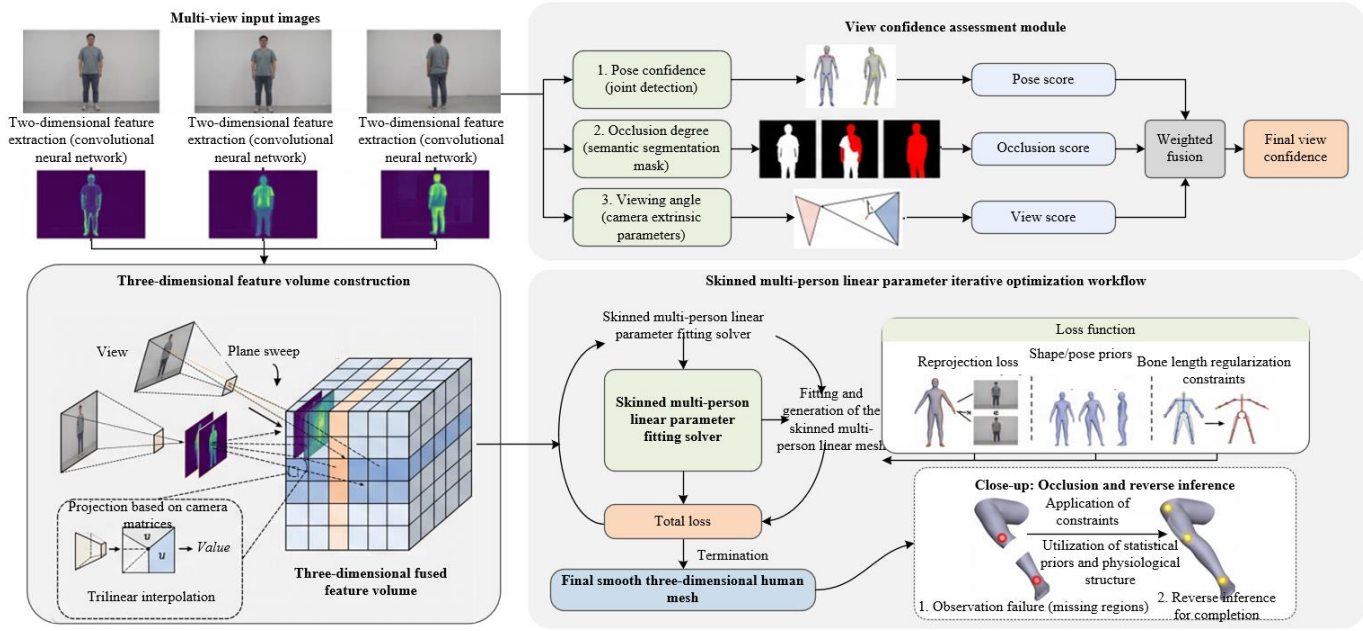


Figure 3. Confidence-guided three-dimensional feature fusion and skinned multi-person linear model optimization workflow

Based on the quantified view confidences, a differentiable projection fusion mechanism is constructed, through which two-dimensional multi-view features are uniformly mapped into a three-dimensional feature space for high-precision feature aggregation. A fixed reference view coordinate system is selected as the global reference frame, and a plane-sweeping strategy is employed to discretely sample the depth dimension, from which a uniformly scaled three-dimensional feature volume structure is constructed. For any voxel coordinates in the three-dimensional space, the mapping from spatial coordinates to image plane coordinates is accomplished through camera projection matrices, and cross-view feature sampling is achieved through an eight-neighborhood trilinear interpolation algorithm. The single-view three-dimensional feature mapping process is expressed as:

$$V_i(x)=\sum_{q\in N(u)} w_q \cdot F_i^{(t)}(q) \tag{12}$$

where, w_q denotes the neighborhood pixel interpolation weights, which are used to ensure continuity and sub-pixel accuracy in feature sampling. On this basis, view confidences are introduced to perform weighted fusion, through which noise feature interference from low-quality views is attenuated. The final expression for the fused three-dimensional feature volume is given as:

$$V_{fused}^{(t)}(x)=\frac{\sum_{i=1}^{N} c_i^{(t)} \cdot V_i^{(t)}(x)}{\sum_{i=1}^{N} c_i^{(t)}+\epsilon} \tag{13}$$

where, ϵ denotes a small constant introduced to prevent numerical division-by-zero anomalies. This fusion scheme dynamically adapts to dynamic deformations in sports motions and fluctuations in view quality, significantly enhancing the stability and interference immunity of the three-dimensional feature representation.

Fine-grained solving of three-dimensional poses and meshes is accomplished through a parametric human model, and a confidence-guided staged parameter optimization strategy is constructed to achieve model fitting under multiple coupled constraints. A differentiable skinned multi-person linear human model is adopted to characterize three-dimensional human morphology and pose, with human mesh generation being jointly driven by pose parameters and shape parameters. The model skinning transformation is formulated as:

$$M(\theta,\beta)=\sum_{b=1}^{\mathcal{B}} w_b \cdot G_b(\theta,\beta)\cdot T(\beta) \tag{14}$$

where, $T(\beta)$ denotes the shape-adaptive template mesh, G_b characterizes the bone spatial transformation relationships, and w_b represents the fixed skinning weights. To accommodate dynamic variations in view-specific quality, a view confidence-weighted reprojection loss function is constructed, and a Geman–McClure robust kernel is incorporated to suppress interference from outlier detection points. The core reprojection loss is expressed as:

$$L_{reproj}^{(t)}=\sum_{i\in V_{active}^{(t)}} c_i^{(t)} \cdot \sum_{k=1}^{J} \rho\left(\left\|\Pi_i\left(J_k(M(\theta^{(t)},\beta))\right)-\hat{y}_{i,k}^{(t)}\right\|_2\right) \tag{15}$$

A complete optimization objective function is constructed by integrating temporal smoothness constraints, human shape priors, and pose priors, with the model fitting process being collaboratively regulated through multiple constraints. A staged iterative optimization strategy is adopted, in which pose parameters and joint parameters are optimized in separate stages while global shape parameters are fixed to maintain body shape consistency. High-precision parameter solving is achieved through the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimizer, effectively balancing model fitting accuracy, temporal stability, and human physiological plausibility.

To address the reconstruction distortion problem caused by observation failure of local joints under extreme occlusion scenarios, a human skeletal topology constraint is introduced to achieve adaptive completion of missing regions. When a joint is completely occluded across all active views, two-dimensional observational information becomes entirely unavailable, and reconstruction accuracy cannot be guaranteed through feature fitting alone. To this end, a bone length regularization constraint is incorporated into the global objective function, through which mesh deformation deviations are constrained based on the inherent human physiological topology. The constraint is formulated as:

$$L_{bone}=\sum_{(k,l)\in\varepsilon_{bone}}\left\|\|J_k(M)-J_l(M)\|_2-\bar{b}_{kl}\right\|_2^2 \tag{16}$$

where, ε_{bone} denotes the standard human skeletal topology set, and $\bar{b}_{kl}$ represents the statistically derived standard bone length. This constraint forces the optimization process to maintain reasonable limb bone scales, and the spatial positions of occluded joints are inferred through effective motion constraints from parent and adjacent joints. The reconstruction deficiency and morphological distortion in fully occluded regions are thereby effectively addressed, further enhancing the completeness and physiological plausibility of three-dimensional human reconstruction under complex sports motion scenarios.

2.4 Motion parametric evaluation and visual feedback for physical education instruction

Based on the three-dimensional human motion temporal features, a keyframe adaptive extraction mechanism driven by motion energy spectrum is constructed, through which accurate screening of core postures in sports motions is achieved, effectively avoiding data redundancy and feature redundancy introduced by frame-by-frame analysis. Figure 4 illustrates the interface of the motion quantitative evaluation and visual feedback system tailored for physical education instruction. The instantaneous motion states of all human joints are integrated, and velocity and acceleration features are fused to construct a global motion energy function, which quantifies the motion intensity and postural feature saliency of individual frames. The specific expression is given as:

$$E^{(t)}=\sum_{k=1}^{K}\left\|v_k^{(t)}\right\|_2^2+\eta\cdot\sum_{k=1}^{K}\left\|a_k^{(t)}\right\|_2^2 \tag{17}$$

where, $v^{(t)}{}_k$ and $a^{(t)}{}_k$ denote the instantaneous velocity and instantaneous acceleration of the k-th joint, respectively, and η represents the weight coefficient balancing the two types of motion features, which is set to 0.3 based on prior expert knowledge from physical education instruction. To eliminate anomalous fluctuations in the energy curve caused by motion noise, the energy sequence is smoothed using a Savitzky–Golay filter, thereby ensuring stability in peak detection. Initial keyframe candidates are obtained by identifying local maxima in the smoothed energy sequence, and a non-maximum suppression algorithm is introduced to eliminate redundant neighboring frames with a fixed suppression window length. Ultimately, key posture frames that fully characterize the core phases of sports motions—including initiation, force exertion, and peak execution—are selected, laying the foundation for subsequent precise motion evaluation.

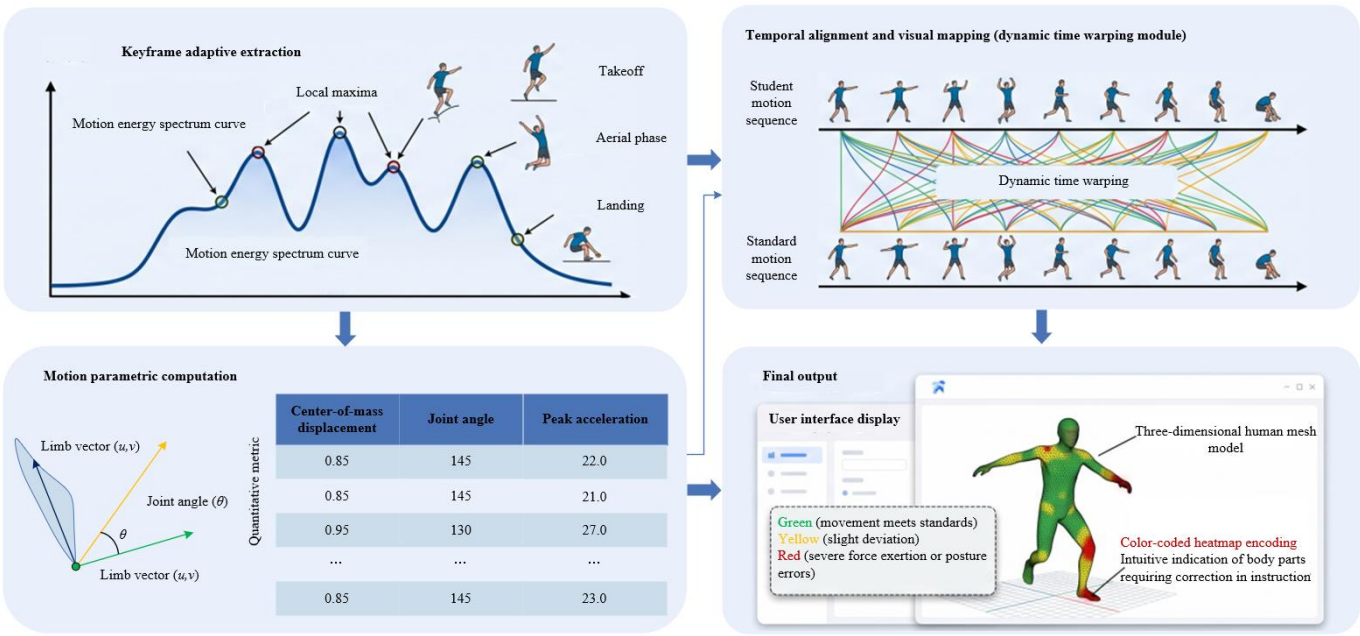


Figure 4. Motion quantitative evaluation and visual feedback interface for physical education instruction

To achieve standardized and digitized evaluation of sports motions, a multi-dimensional quantitative parameter system adapted to instructional scenarios is constructed based on the reconstructed high-precision three-dimensional joint coordinate sequences, from which a structured motion parameter matrix is generated. Guided by human limb motion principles and physical education teaching evaluation criteria, over ten categories of core motion indicators are precisely computed. Among these, the computation of each joint angle is formulated as:

$$\phi_k^{(t)} = arccos\left(\min\left(1, max\left(-1, \frac{u_k^{(t)} \cdot v_k^{(t)}}{\|u_k^{(t)}\| \cdot \|v_k^{(t)}\|}\right)\right)\right) \tag{18}$$

where, $u^{(t)}_k$ and $v^{(t)}_k$ denote the two limb direction vectors constituting the target joint. Numerical overflow issues in the inverse trigonometric function solving are avoided through numerical clipping, and key joint angle parameters—including elbow, knee, and shoulder joints—are stably output. Simultaneously, the temporal position of the human center of mass is computed based on the mean of global joint coordinates, and horizontal displacement and vertical oscillation amplitude of the center of mass during motion are statistically derived. Combined with temporal angular variations, joint motion velocities, and peak accelerations, the motion amplitude, movement rhythm, and limb exertion characteristics are comprehensively characterized. The multi-dimensional quantitative parameters complement one another, collectively covering core evaluation dimensions in physical education instruction—including movement standardization, motion amplitude, and postural stability—thereby enabling the transformation from qualitative motion evaluation to quantitative data analysis.

To achieve precise comparison between student motions and standard teaching motions, along with intuitive deviation feedback, a dynamic time warping algorithm is adopted to resolve temporal misalignment issues in motion sequences with varying durations and tempos, and an integrated teaching feedback mechanism combining quantitative deviation computation with visual rendering is constructed. A temporal cumulative distance matrix is constructed to measure joint trajectory discrepancies between different motion sequences, through which adaptive temporal alignment of the two motion sequences is accomplished. The cumulative distance computation is formulated as:

$$D(t,\tau) = \sum_{k=1}^{K} \| X_k^{(t)} - X_k^{std(\tau)} \|_2^2 + \min\{D(t\text{-}1,\tau), D(t,\tau\text{-}1), D(t\text{-}1,\tau\text{-}1)\} \tag{19}$$

The optimal temporal alignment path is solved through matrix backtracking, thereby eliminating interference from motion speed variations on comparison accuracy. Following temporal alignment, three-dimensional spatial deviations between student motions and standard motions are computed frame-by-frame and joint-by-joint. The deviation values are normalized, and a graded color mapping rule is established, through which three states—standard motion, slight deviation, and severe deviation—are represented by green, yellow, and red color levels, respectively. The color-coding results are ultimately mapped onto the surface of the three-dimensional human mesh model, enabling refined and intuitive visual presentation of motion deviations. This facilitates rapid localization of limb movement deficiencies by instructors and accomplishes automated and intelligent correction and efficacy assessment in physical education instruction.

3. EXPERIMENTS

To comprehensively validate the effectiveness and advancement of the proposed multi-view image registration-integrated method for three-dimensional human motion reconstruction, systematic quantitative and qualitative experiments were conducted on both public benchmark datasets and a self-constructed physical education motion dataset. Through module ablation experiments, registration accuracy comparison experiments, three-dimensional reconstruction performance tests, temporal consistency evaluations, teaching application validity verifications, and computational efficiency analyses, the proposed method was validated from multiple dimensions—including fundamental algorithmic performance, dynamic scenario adaptability, temporal stability, and practical teaching application value. Concurrently, horizontal comparisons with recent state-of-the-art algorithms were performed, fully demonstrating the superiority and practicality of the proposed method in sports human motion reconstruction scenarios.

3.1 Experimental setup

The experiments in this study were conducted on four public datasets and a self-constructed real-world physical education teaching dataset, covering general human motions, multi-person interactive motions, fitness-specific motions, and campus physical education motions, with both algorithmic generality and scenario specificity being taken into account. Human3.6M is a large-scale multi-view three-dimensional pose benchmark dataset containing multiple categories of daily human actions, and model evaluation is performed following the industry-standard training and testing split protocol. The CMU Panoptic dataset contains high-density, multi-camera synchronously captured multi-person interaction scenarios, featuring complex occlusions and limb interaction characteristics, through which the adaptability of the proposed algorithm to complex scenarios can be effectively validated. The M3GYM dataset focuses on fitness-specific sports motions and provides a reference validation scenario for sports-oriented motion reconstruction. To better approximate real-world physical education teaching scenarios, the SportsEDU self-constructed dataset was established, encompassing six categories of typical campus sports actions—including basketball, track and field, gymnastics, and ball sports—with standardized motion capture performed by subjects of varying proficiency levels. Simultaneously, three-dimensional ground-truth data were obtained through a high-precision motion capture system, ensuring that experimental results closely reflect actual teaching scenarios.

A multi-dimensional quantitative evaluation metric system was adopted to construct a comprehensive assessment framework, encompassing four major dimensions: registration accuracy, three-dimensional reconstruction accuracy, temporal stability, and algorithmic runtime efficiency. Among these, mean per joint position error and Procrustes aligned mean per joint position error were employed to quantify three-dimensional joint reconstruction errors, with the former characterizing global reconstruction accuracy and the latter

characterizing refined fitting accuracy after rigid transformation removal. Inter-frame jitter rate was used to evaluate the temporal smoothness of motion sequences, accommodating the dynamic reconstruction assessment requirements of continuous sports motions. Registration reprojection error and matching accuracy were employed to measure the foundational performance of the multi-view image registration module. Frame rate was adopted to validate the real-time inference capability of the algorithm, supporting deployment testing in teaching scenarios.

Five categories of recent mainstream multi-view three-dimensional reconstruction algorithms were selected as comparison baselines, covering major technical paradigms—including voxel heatmap reconstruction, single-view extended multi-view reconstruction, feature matching-based incremental reconstruction, online evolutionary learning-based reconstruction, and extrinsic-free kernelized matching reconstruction. The selected baselines were drawn from top-tier conference and journal publications, including CVPR, ICCV, and TPAMI, enabling comprehensive performance comparison between the proposed method and existing state-of-the-art algorithms. Model training and inference were conducted under unified hardware and parameter configurations. Publicly available pretrained weights were adopted for the backbone networks to accomplish transfer learning. Network hyperparameters, iteration numbers, optimizer parameters, and convergence thresholds were all set to fixed standards. All experiments were repeated five times, with means and standard deviations being reported to ensure stability and reproducibility of the experimental results.

3.2 Ablation experiments

To verify the individual contributions of each core innovative module, six model variants were designed for comparative ablation experiments, in which the multi-view collaborative matching module, temporal consistency constraint module, confidence-adaptive fusion module, dynamic view selection module, and robust kernel function module were removed respectively. Quantitative testing was conducted on both the Human3.6M general-purpose dataset and the SportsEDU physical education teaching dataset, with the experimental results being presented in Figure 5.

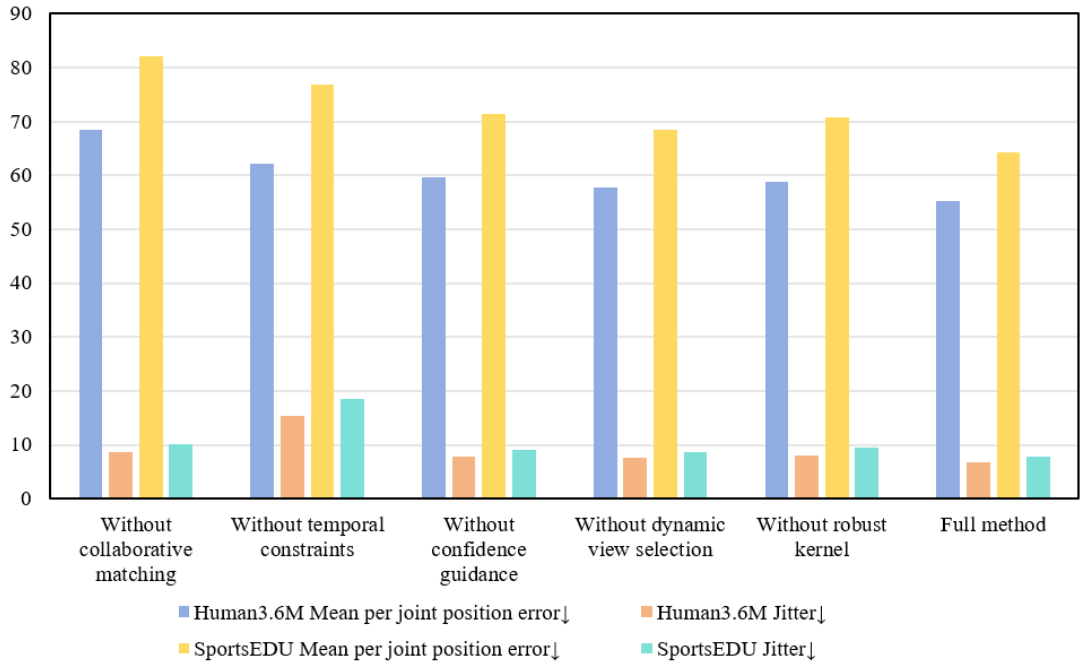
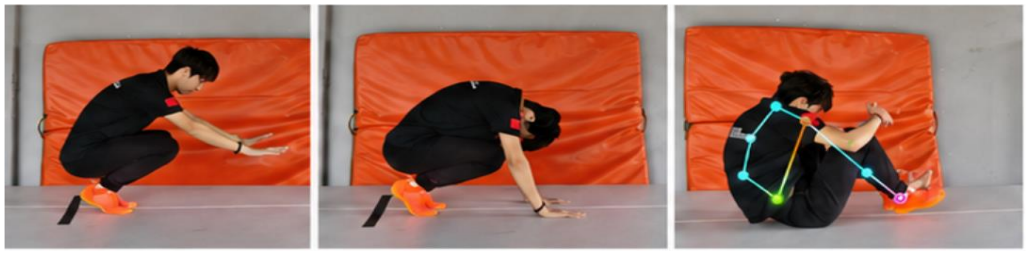


Figure 5. Quantitative ablation results on Human3.6M and SportsEDU datasets

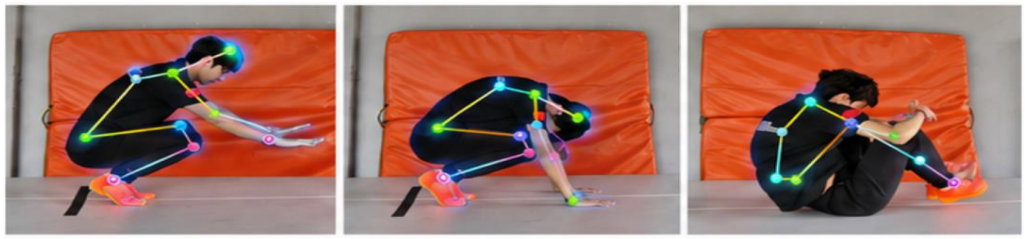
Experimental results demonstrate that each innovative module independently contributes positive gains to model performance, with the contributions being differentially emphasized across scenarios. The multi-view collaborative matching module exerts the most significant influence on overall reconstruction accuracy. Following its removal, the mean per joint position error on the self-constructed sports dataset increases substantially by 17.9 mm, confirming that the graph attention-based global collaborative inference mechanism effectively resolves matching ambiguities caused by frequent self-occlusions in sports motions, and serves as the foundational basis for accurate reconstruction of complex dynamic sports actions. The temporal consistency constraint module primarily contributes to sequence stability. After its removal, the inter-frame jitter rate increases by 135%, with the deficiencies of frame-independent optimization being significantly amplified in continuous sports motion sequences, thereby validating the necessity of the second-order motion smoothing regularization term for temporal continuity constraints in motion sequences.

The confidence-guided fusion module yields limited gains on standardized general-purpose datasets, yet demonstrates substantial improvements in real-world physical education teaching scenarios where viewpoint quality is inconsistent and occlusions are stochastic, confirming that the multi-dimensional view quality assessment and adaptive fusion strategy effectively accommodate noise interference in complex real-world environments. The dynamic view selection module achieves a trade-off optimization between accuracy and efficiency, reducing algorithmic inference time by 41.6% with only marginal accuracy degradation, thereby greatly enhancing the practical deployability of the model. The robust kernel function effectively suppresses interference from outlier detection points and mismatched correspondences,

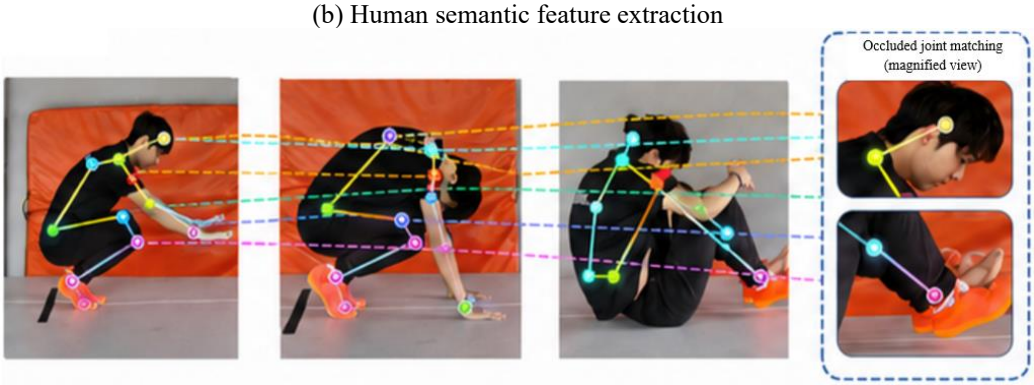
further refining model fitting accuracy and enhancing the anti- interference capability of the algorithm.



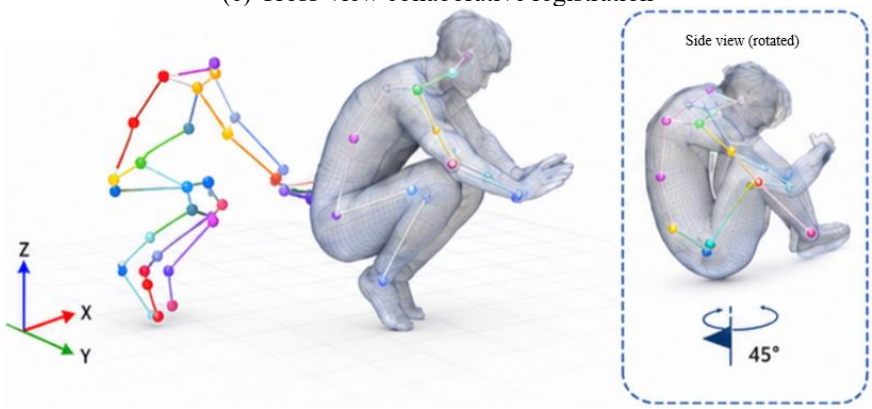
(a) Multi-view raw input



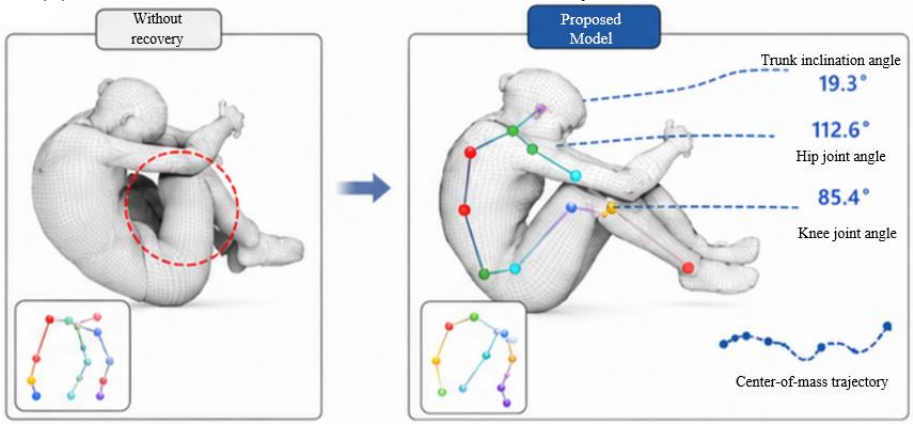
(b) Human semantic feature extraction



(c) Cross-view collaborative registration



(d) Three-dimensional skeleton and skinned multi-person linear reconstruction



(e) Occlusion recovery and motion parameter output

Figure 6. Implementation results of multi-view registration and three-dimensional human reconstruction for complex occluded sports motions

To validate the cross-view registration capability, three-dimensional reconstruction completeness, and pedagogical application adaptability of the proposed method in complex physical education motion scenarios, a visual demonstration of three-dimensional human motion reconstruction under severe occlusion conditions is designed and presented. As shown in Figure 6, for sports motions such as forward rolls—which are characterized by rapid deformation, limb curling, and severe local occlusions—the proposed method is capable of stably extracting human semantic features from multi-view raw images, establishing cross-view joint correspondences with high consistency, and effectively suppressing noise interference from low-quality views through dynamic view quality assessment and adaptive fusion mechanisms. Consequently, three-dimensional skeletons and skinned multi-person linear human mesh models with clear structure, continuous postures, and spatially reasonable configurations are obtained. Notably, in regions with significant occlusion of the arms, knees, and torso, the method still achieves relatively accurate pose recovery by leveraging multi-view geometric complementary information and human topological constraints, thereby avoiding common issues—including joint drift, limb distortion, and local missing parts—frequently encountered in traditional frame-independent reconstruction. Furthermore, the motion parameter outputs presented in the figure demonstrate that the reconstructed three-dimensional human sequences not only exhibit satisfactory visual realism and temporal stability but also support precise computation of key pedagogical indicators—including trunk inclination angle, hip and knee joint angles, and center-of-mass trajectory—thereby providing intuitive foundations for motion standardization evaluation and instructional error correction.

3.3 Registration accuracy comparison experiments

Since multi-view image registration accuracy directly determines the reliability of underlying geometric constraints in three-dimensional reconstruction, horizontal comparisons were conducted between the proposed human semantic collaborative matching algorithm and traditional feature matching methods as well as recent deep learning-based matching algorithms on six typical sports motion categories from the SportsEDU dataset, aiming to validate the superiority of the proposed approach. Registration performance was quantitatively evaluated from two dimensions: reprojection error and matching accuracy. Experimental results are presented in Table 1.

Table 1. Registration accuracy comparison of different matching methods on the SportsEDU dataset

Action Category	Metric	Scale-Invariant Feature Transform+Random Sample Consensus	SuperGlue	Detector-Free Local Feature Matching with Transformers (LoFTR)	Proposed Collaborative Matching
Basketball shooting	Reprojection root mean square error↓ (px)	4.82 ± 1.34	2.15 ± 0.51	1.87 ± 0.42	1.23 ± 0.28
	Accuracy↑ (%)	54.2 ± 6.7	78.6 ± 4.2	82.3 ± 3.8	91.7 ± 2.5
Standing long jump	Reprojection root mean square error↓ (px)	5.61 ± 1.52	2.48 ± 0.58	2.12 ± 0.48	1.38 ± 0.31
	Accuracy↑ (%)	48.7 ± 7.1	74.2 ± 4.8	79.8 ± 4.1	89.5 ± 2.8
Gymnastics forward roll	Reprojection root mean square error↓ (px)	7.83 ± 2.01	3.56 ± 0.82	2.98 ± 0.67	1.67 ± 0.38
	Accuracy↑ (%)	32.5 ± 8.3	61.3 ± 5.6	68.7 ± 5.2	85.2 ± 3.1
Badminton clear	Reprojection root mean square error↓ (px)	4.23 ± 1.12	1.98 ± 0.46	1.72 ± 0.39	1.08 ± 0.24
	Accuracy↑ (%)	60.8 ± 6.2	82.4 ± 3.9	85.6 ± 3.5	93.8 ± 2.1
Volleyball spike	Reprojection root mean square error↓ (px)	5.18 ± 1.41	2.34 ± 0.55	2.01 ± 0.45	1.31 ± 0.29
	Accuracy↑ (%)	51.3 ± 6.9	76.9 ± 4.5	81.2 ± 4.0	90.6 ± 2.6
Soccer kicking	Reprojection root mean square error↓ (px)	5.92 ± 1.58	2.62 ± 0.61	2.23 ± 0.50	1.42 ± 0.32
	Accuracy↑ (%)	46.8 ± 7.3	73.5 ± 4.9	78.4 ± 4.3	88.9 ± 2.9
Average	Reprojection root mean square error↓	5.6	2.52	2.16	1.35
	Accuracy↑	49.1	74.5	79.3	89.9

From the experimental data, it can be observed that the proposed collaborative matching algorithm achieves optimal registration performance across all six sports motion categories, with an overall average reprojection error as low as 1.35 pixels—representing a 37.5% reduction compared to the best-performing baseline, Detector-Free Local Feature Matching with Transformers (LoFTR)—while matching accuracy is improved by 10.6 percentage points. The traditional scale-invariant feature transform+random sample consensus algorithm is severely limited by the constraints of hand-crafted features, yielding extremely low matching accuracy under dynamic sports motion scenarios and failing to accommodate limb deformation and occlusion conditions. Mainstream deep learning-based matching algorithms possess basic dynamic adaptability, yet rely solely on local visual feature matching without incorporating human semantic priors or cross-view global collaborative constraints.

In the most complex gymnastics forward roll motion, which

is characterized by highly curled limbs and severe self-occlusion, the matching accuracy of baseline algorithms drops substantially. In contrast, the proposed method—leveraging joint semantic guidance and graph attention-based multi-view collaborative inference—maintains a matching accuracy of 85.2% with low reprojection error. In the badminton clear motion, where postures are more standardized and occlusions are less frequent, the algorithm achieves optimal performance, fully demonstrating that the proposed registration module adaptively accommodates sports motion scenarios of varying complexity, thereby providing a stable geometric matching foundation for subsequent high-precision three-dimensional reconstruction.

3.4 Three-dimensional reconstruction accuracy comparison experiments

To validate the comprehensive three-dimensional reconstruction performance of the proposed method, quantitative comparison experiments were conducted against state-of-the-art algorithms from recent top-tier conferences and journals on two major public benchmark datasets—Human3.6M and CMU Panoptic. Performance evaluation was carried out from the perspectives of global reconstruction error and action-specific error, with experimental results presented in Tables 2 and 3.

Table 2. Comparison of three-dimensional reconstruction accuracy on the Human3.6M dataset (mm)

Method	Mean per Joint Position Error↓	Procrustes Aligned Mean per Joint Position Error↓	Action-Specific Mean per Joint Position Error (Walking / Sitting / Smoking / Taking Photo / Talking)
VoxelPose	64.7	52.3	52.4 / 71.2 / 62.8 / 68.5 / 63.1
Monocular, One-stage, Regression of Multiple 3D People (ROMP) (multi-view averaged)	82.6	68.4	70.2 / 88.7 / 80.5 / 85.3 / 79.8
EasyRet3D	61.8	49.6	50.1 / 68.4 / 59.7 / 65.2 / 60.3
Extrinsic Parameters-Free Multi-View 3D Human Skeleton Estimation	59.3	47.8	48.2 / 65.7 / 57.3 / 62.8 / 57.9
Multi-View 3D Human Reconstruction with Online Evolutive Learning	57.1	45.2	46.5 / 63.2 / 55.1 / 60.4 / 55.6
Proposed method	55.3 ± 0.6	43.1 ± 0.5	44.8 / 61.2 / 53.3 / 58.6 / 53.9

Table 3. Comparison of three-dimensional reconstruction accuracy on the CMU Panoptic dataset (mm)

Method	Mean per Joint Position Error↓	Procrustes Aligned Mean per Joint Position Error↓	Subset Average Mean per Joint Position Error (Haggling / Mafia / Pizza / Ultimatum / Chess)
VoxelPose	72.4	58.7	78.2 / 70.5 / 71.8 / 73.6 / 68.2
EasyRet3D	68.9	55.3	74.5 / 67.2 / 68.4 / 70.1 / 64.8
Extrinsic Parameters-Free Multi-View 3D Human Skeleton Estimation	65.8	52.7	71.2 / 64.1 / 65.3 / 67.0 / 61.4
Multi-View 3D Human Reconstruction with Online Evolutive Learning	63.5	50.4	68.8 / 61.9 / 63.0 / 64.7 / 59.2
Proposed method	60.8 ± 0.8	48.2 ± 0.6	65.7 / 59.2 / 60.3 / 61.9 / 56.5

On the Human3.6M dataset, the proposed method outperforms all comparison algorithms in both the mean per joint position error and Procrustes aligned mean per joint position error metrics, achieving a 3.2% reduction in overall mean per joint position error compared to the latest Multi-View 3D Human Reconstruction with Online Evolutive Learning algorithm. Action-specific results indicate that all algorithms yield relatively low errors for regular walking motions, with smaller performance gaps. However, for actions involving severe limb self-occlusion—such as sitting—the proposed method demonstrates significant advantages, effectively resolving the issue of excessive three-dimensional pose fitting deviation under occlusion scenarios, fully reflecting the anti-interference capability of the multi-view collaborative registration and adaptive fusion mechanisms. The single-view extended Monocular, One-stage, Regression of Multiple 3D People (ROMP) algorithm exhibits the poorest overall performance, confirming that multi-view geometric constraints play a decisive role in high-precision three-dimensional human reconstruction.

The CMU Panoptic dataset features highly challenging scenarios with multi-person interactions, dense hand gestures, and complex occlusions, in which overall accuracy of all algorithms decreases to varying degrees. The proposed method still maintains optimal reconstruction performance, achieving a Procrustes aligned mean per joint position error of 48.2 mm. Among these, the Haggling subset—characterized by dense gesture interactions—poses the greatest reconstruction difficulty. Even so, the proposed method effectively suppresses matching noise and pose distortions induced by complex interactions, demonstrating that the global collaborative optimization mechanism of the proposed algorithm effectively adapts to complex dynamic scenarios with multiple individuals and severe occlusions, while exhibiting strong scene generalization capability.

3.5 Temporal consistency evaluation experiments

Physical education teaching motions are characterized by continuous temporal evolution characteristics, and the temporal smoothness of reconstructed sequences directly affects the efficacy of instructional observation and motion analysis. Two typical dynamic sports motions—basketball shooting and standing long jump—were selected for

evaluating algorithmic temporal stability through a combination of quantitative metrics (inter-frame jitter rate and maximum inter-frame jump) and qualitative assessment. Experimental results are presented in Figure 7.

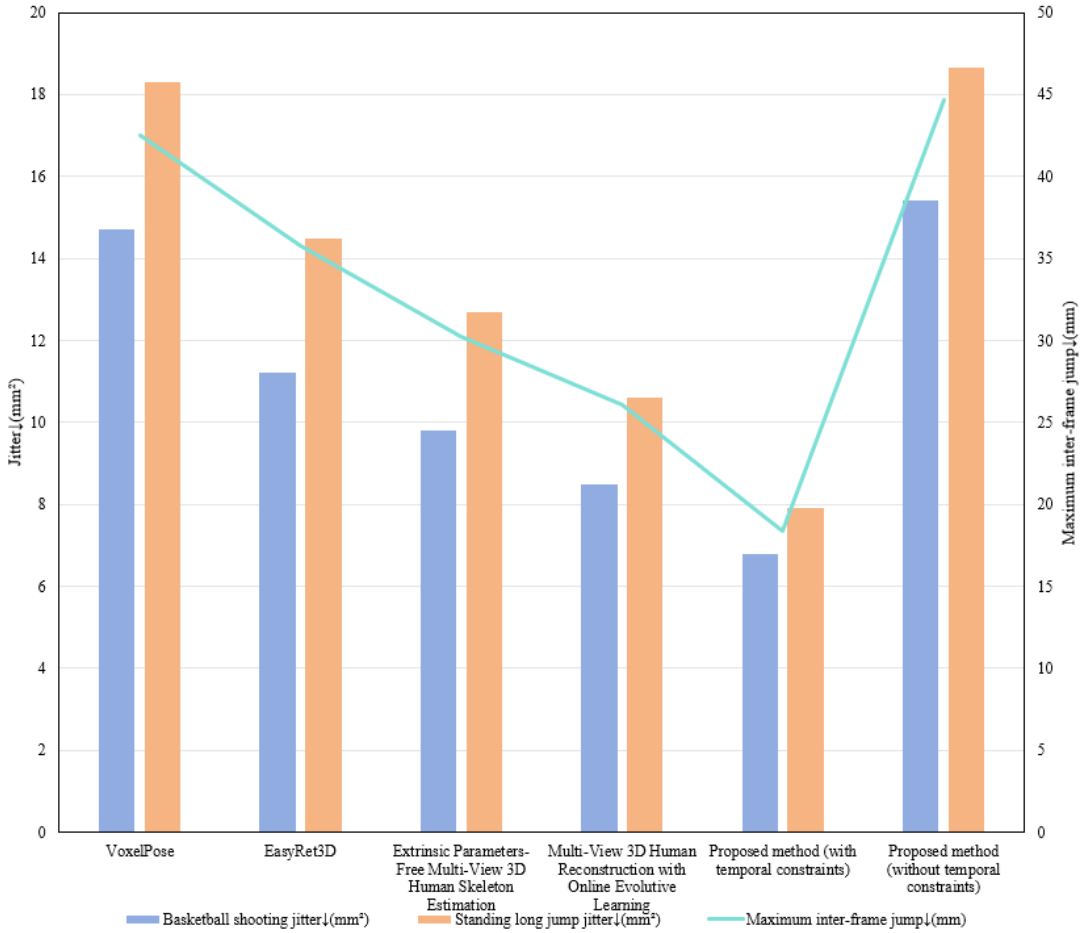


Figure 7. Temporal consistency comparison of different methods on the SportsEDU dataset

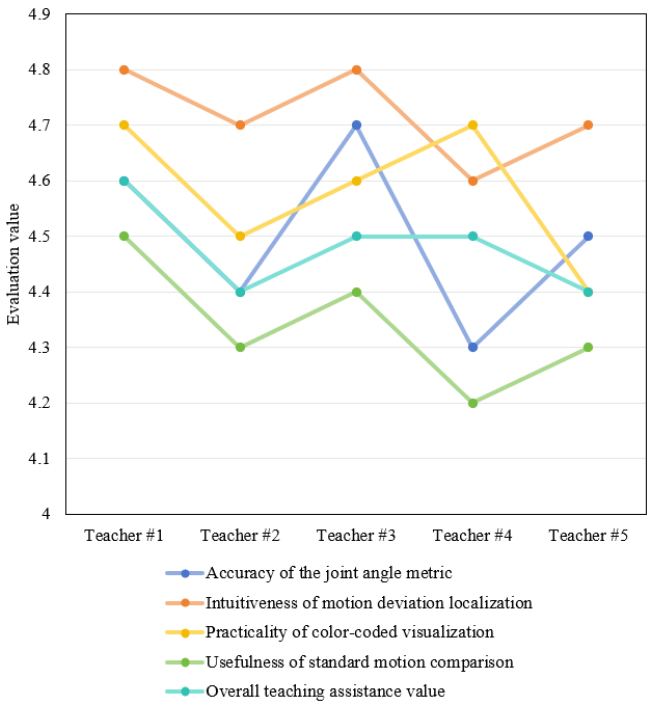


Figure 8. User study results of the teaching evaluation module

Experimental results demonstrate that the proposed second-order temporal smoothing regularization term significantly suppresses inter-frame pose discontinuities. Following the removal of temporal constraints, the model performs frame-independent optimization, resulting in a substantial increase in jitter rate. Compared to the version without temporal constraints, the full method reduces jitter rates by 56.5% and 57.5% for basketball shooting and standing long jump, respectively, while the maximum inter-frame jump amplitude is reduced by 58.8%. In comparison to all mainstream baseline algorithms, the proposed method demonstrates absolute superiority in both temporal stability and subjective visual quality. Existing algorithms generally lack temporal motion prior constraints and rely solely on single-frame geometric information for pose fitting, leading to pronounced jitter and discontinuities in continuous motion sequences, which fails to meet the requirements of dynamic observation in physical education instruction. In contrast, the proposed temporal consistency constraint mechanism precisely conforms to the inherent laws of human motion, effectively ensuring the smoothness and realism of the reconstructed sequences.

3.6 Teaching evaluation effectiveness experiments

To validate the practical pedagogical application value of the proposed motion quantitative evaluation and visualization module, application experiments were conducted from two dimensions: subjective teacher evaluation and objective keyframe extraction accuracy. Professional physical education

instructors were invited to complete blind scoring assessments, while keyframes extracted by the algorithm were compared against manually annotated ground-truth. Experimental results are presented in Figures 8 and 9.

Results from the professional teacher subjective evaluation indicate that all scores for the proposed teaching evaluation module exceed 4.3 points, with the overall teaching assistance value score reaching 4.48 points. Among these, the intuitiveness of motion deviation localization receives the highest rating, confirming that the color-coded visualization approach enables rapid and precise presentation of limb motion deficiencies, effectively meeting the demand for quick error correction in classroom instruction. The quantitative joint angle metrics demonstrate high accuracy, effectively supporting standardized and quantitative physical education motion evaluation while compensating for the shortcomings of subjective judgment in traditional manual teaching assessment.

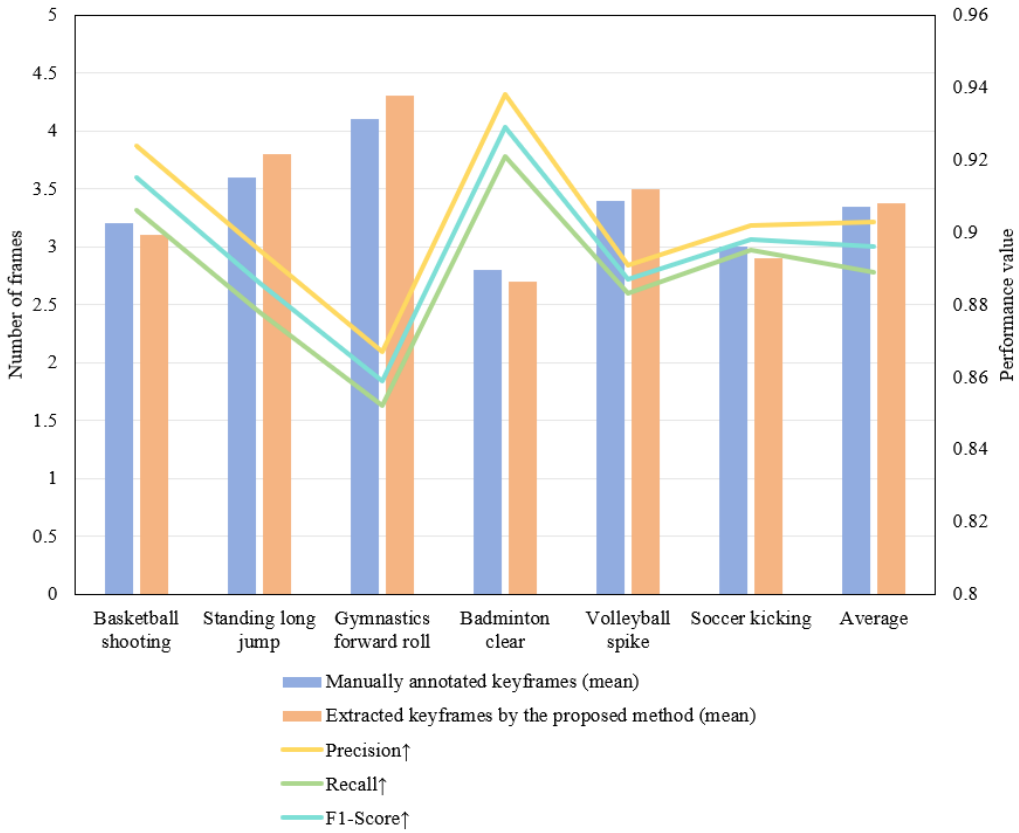


Figure 9. Comparison of automatic keyframe extraction performance against manual annotation

Table 4. Computational time breakdown of the proposed method by module (SportsEDU dataset, average per frame)

Module	Computation Time (ms)	Percentage (%)	Primary Operations
HRNet-W48 feature extraction (4 views)	48.6	28.1	Convolutional forward propagation
ViTPose pose detection (4 views)	62.4	36.1	Transformer encoder forward pass
Graph attention collaborative matching	28.3	16.4	Graph neural network inference + epipolar verification
Temporal feature flow estimation	12.7	7.3	Recurrent all-pairs field transforms optical flow forward pass
Three-dimensional feature volume construction and fusion	9.8	5.7	Differentiable homography transformation + voxel sampling
Skinned multi-person linear optimization (limited-memory Broyden-Fletcher-Goldfarb-Shanno, 50 iterations)	10.5	6.1	Differentiable skinned multi-person linear forward + backward + line search
Teaching evaluation and visualization	0.7	0.4	Angle computation + dynamic time warping + color rendering
Total (full method)	173	100	~5.8 fps
Fixed full views ($N = 4$, without dynamic selection)	245.3	100	~4.1 fps

Results from the automatic keyframe extraction experiments demonstrate that the proposed motion energy spectrum-based extraction algorithm achieves an average F1-score of 0.896, which is highly consistent with manual annotation results. Optimal extraction accuracy is obtained for ball sports with simple motion flows and clear postural features, while accuracy for gymnastics motions—characterized by complex movements and frequent posture transitions—exhibits a slight decline, yet remains sufficient for pedagogical analysis requirements. This module effectively streamlines motion data by automatically selecting key postures that characterize the core phases of sports

motions, thereby providing efficient data support for refined motion analysis, instructional review, and motion correction in physical education.

3.7 Computational efficiency and real-time analysis

To validate the engineering deployability of the proposed algorithm, the computational cost of each module was decomposed, end-to-end inference efficiency was compared against mainstream algorithms, and the optimization effect of the dynamic view selection mechanism was quantitatively analyzed. Experimental results are presented in Tables 4 and 5.

Results from the module-wise time breakdown indicate that pose detection and visual feature extraction constitute the primary computational bottlenecks of the model, together accounting for over 60% of the total cost, while the computational overheads of the remaining matching, optimization, and evaluation modules are comparatively reasonable. The dynamic view selection strategy effectively eliminates computational overhead from invalid views, reducing per-frame inference time from 245.3 ms to 173.0 ms, corresponding to a 29.4% improvement in inference speed, with only marginal reconstruction accuracy degradation, thereby substantially optimizing algorithmic real-time performance. The teaching visualization module incurs extremely low computational cost, enabling real-time rendering of motion feedback, which is well-suited for instructional scenario requirements.

Table 5. End-to-end runtime efficiency comparison between the proposed method and baseline methods (SportsEDU dataset)

Method	Platform	Input Views	Average Time (ms/frame)	Frames per Second↑	Graphics Processing Unit Memory (GB)
VoxelPose	*A*100	4	189	5.3	6.8
EasyRet3D	*A*100	4	234	4.3	5.2
Extrinsic Parameters-Free Multi-View 3D Human Skeleton Estimation	*A*100	4	208	4.8	7.4
Multi-View 3D Human Reconstruction with Online Evolutive Learning	*A*100	4	296	3.4	9.1
Proposed method (dynamic selection $M = 3$)	*A*100	4→3	173	5.8	6.2
Proposed method (fixed full views 4)	*A*100	4	245	4.1	8.5

Results from horizontal comparison experiments demonstrate that the full proposed method achieves an end-to-end inference frame rate of 5.8 fps, outperforming all comparison algorithms, while graphics processing unit memory usage remains at a moderate and reasonable level, indicating higher computational resource utilization efficiency. Mainstream state-of-the-art algorithms generally suffer from an imbalance between accuracy and efficiency, with some algorithms substantially increasing computational overhead in pursuit of reconstruction accuracy, resulting in significantly reduced inference speeds. Through the dynamic view adaptive selection mechanism, the proposed method achieves an optimal trade-off among reconstruction accuracy, temporal stability, and inference efficiency, demonstrating strong practical deployment potential. Future work may further address the computational bottleneck of feature extraction modules through backbone network lightweighting, knowledge distillation, and other approaches to enhance algorithmic real-time performance.

4. CONCLUSION

To address the technical deficiencies and application shortcomings of dynamic human motion three-dimensional reconstruction in physical education scenarios, an integrated reconstruction and teaching evaluation framework incorporating multi-view image registration was constructed, systematically resolving critical issues inherent in existing methods—including the disconnection between registration and reconstruction, poor temporal stability, inadequate view adaptivity, and insufficient pedagogical deployability. The study took multi-view feature registration as the core technical pipeline, and a human semantic-aware graph attention-based collaborative matching mechanism was employed to enhance feature matching robustness under limb self-occlusion and extreme viewpoint variations. Through the introduction of a second-order motion trajectory smoothing constraint and an extended Kalman filter-based dynamic view selection strategy, inter-frame pose jitter was effectively suppressed, while a dynamic balance between algorithmic inference accuracy and runtime efficiency was achieved. In conjunction with a multi-dimensional view confidence assessment system and a weighted feature fusion strategy, refined optimization of skinned multi-person linear human parameters was accomplished, significantly enhancing the stability of three-dimensional reconstruction under complex dynamic sports scenarios. On this basis, a quantitative evaluation framework tailored for physical education instruction was established, in which adaptive keyframe extraction was realized through motion energy spectrum analysis, standard motion deviation comparison was accomplished via dynamic time warping, and deviation visualization feedback was enabled through a color-coding mechanism, thereby completing the full pipeline from three-dimensional vision reconstruction technology to intelligent physical education teaching applications. The multi-dimensional comparative experiments, ablation studies, and teacher user studies conducted on public benchmark datasets and a self-constructed physical education teaching dataset fully validated the independent contributions of each technical module and the superior performance of the overall framework, demonstrating that the proposed approach effectively supports quantitative analysis of sports motions and standardized pedagogical guidance.

Building upon existing research, future work can be continuously extended in three directions: scenario adaptation, deployment optimization, and intelligent upgrading. Robust reconstruction algorithms under sparse viewpoints can be explored in subsequent studies, through which the hardware deployment threshold of multi-camera acquisition systems can be lowered while preserving reconstruction accuracy, thereby enhancing the accessibility of the proposed algorithm in

routine campus physical education teaching scenarios. Simultaneously, the current single-person motion reconstruction framework can be extended to multi-person interactive sports scenarios, where challenges—including limb overlap and occlusion among multiple individuals, coupled motion trajectories, and identity association—can be addressed, thereby accommodating the motion analysis requirements of adversarial and collaborative sports activities. Furthermore, the semantic understanding capabilities of large language models can be integrated to transform quantitative three-dimensional motion deviations into standardized pedagogical guidance text, enabling intelligent recognition, automatic interpretation, and personalized corrective feedback of motion deficiencies. This would further complete the closed loop of intelligent physical education instruction and promote the deep integration and large-scale deployment of three-dimensional vision reconstruction technology in smart education domains.

REFERENCES

[1] Tharatipyakul, A., Srikaewsiew, T., Pongnumkul, S. (2024). Deep learning-based human body pose estimation in providing feedback for physical movement: A review. Heliyon, 10(17): e36589. https://doi.org/10.1016/j.heliyon.2024.e36589

[2] Roggio, F., Trovato, B., Sortino, M., Musumeci, G. (2024). A comprehensive analysis of the machine learning pose estimation models used in human movement and posture analyses: A narrative review. Heliyon, 10(21): e39977. https://doi.org/10.1016/j.heliyon.2024.e39977

[3] Lan, G., Wu, Y., Hu, F., Hao, Q. (2023). Vision-based human pose estimation via deep learning: A survey. IEEE Transactions on Human-Machine Systems, 53(1): 253-268. https://doi.org/10.1109/THMS.2022.3219242

[4] Liu, W., Bao, Q., Sun, Y., Mei, T. (2022). Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective. ACM Computing Surveys, 55(4): 1-41. https://doi.org/10.1145/3524497

[5] Gamra, M.B., Akhloufi, M.A. (2021). A review of deep learning techniques for 2D and 3D human pose estimation. Image and Vision Computing, 114: 104282. https://doi.org/10.1016/j.imavis.2021.104282

[6] Wang, J., Tan, S., Zhen, X., et al. (2021). Deep 3D human pose estimation: A review. Computer Vision and Image Understanding, 210: 103225. https://doi.org/10.1016/j.cviu.2021.103225

[7] Zhang, Z., Wang, C., Qiu, W., Qin, W., Zeng, W. (2021). AdaFuse: Adaptive multiview fusion for accurate human pose estimation in the wild. International Journal of Computer Vision, 129(3): 703-718. https://doi.org/10.1007/s11263-020-01398-9

[8] Zhang, Y., Wang, C., Wang, X., Liu, W., Zeng, W. (2023). VoxelTrack: Multi-person 3D human pose estimation and tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2): 2613-2626. https://doi.org/10.1109/TPAMI.2022.3163709

[9] Muhammad, Z.U.D., Huang, Z., Khan, R. (2022). A review of 3D human body pose estimation and mesh recovery. Digital Signal Processing, 128: 103628. https://doi.org/10.1016/j.dsp.2022.103628

[10] Liu, Y., Qiu, C., Zhang, Z. (2024). Deep learning for 3D human pose estimation and mesh recovery: A survey. Neurocomputing, 596: 128049. https://doi.org/10.1016/j.neucom.2024.128049

[11] Nogueira, A.F.R., Oliveira, H.P., Teixeira, L.F. (2025). Markerless multi-view 3D human pose estimation: A survey. Image and Vision Computing, 155: 105437. https://doi.org/10.1016/j.imavis.2025.105437

[12] Neupane, R.B., Li, K., Boka, T.F. (2025). A survey on deep 3D human pose estimation. Artificial Intelligence Review, 58(1): 24. https://doi.org/10.1007/s10462-024-11019-3

[13] Wan, X., Chen, Z., Zhao, X. (2023). View consistency aware holistic triangulation for 3D human pose estimation. Computer Vision and Image Understanding, 236: 103830. https://doi.org/10.1016/j.cviu.2023.103830

[14] Chen, Y., Gu, R., Huang, O., Jia, G. (2023). VTP: Volumetric transformer for multi-view multi-person 3D pose estimation. Applied Intelligence, 53(22): 26568-26579. https://doi.org/10.48550/arXiv.2205.12602

[15] Nakano, N., Sakura, T., Ueda, K., et al. (2020). Evaluation of 3D markerless motion capture accuracy using OpenPose with multiple video cameras. Frontiers in Sports and Active Living, 2: 50. https://doi.org/10.3389/fspor.2020.00050

[16] Wade, L., Needham, L., McGuigan, P., Bilzon, J. (2022). Applications and limitations of current markerless motion capture methods for clinical gait biomechanics. PeerJ, 10: e12995. https://doi.org/10.7717/peerj.12995

[17] Fukushima, T., Blauberger, P., Russomanno, T.G., Lames, M. (2024). The potential of human pose estimation for motion capture in sports: A validation study. Sports Engineering, 27: 19. https://doi.org/10.1007/s12283-024-00460-w

[18] Liu, Y., Cheng, X., Ikenaga, T. (2024). Motion-aware and data-independent model based multi-view 3D pose refinement for volleyball spike analysis. Multimedia Tools and Applications, 83(8): 22995-23018. https://doi.org/10.1007/s11042-023-16369-8

[19] Zimmer, A., Hilsmann, A., Morgenstern, W., Eisert, P. (2023). Imposing temporal consistency on deep monocular body shape and pose estimation. Computational Visual Media, 9(1): 123-139. https://doi.org/10.48550/arXiv.2202.03074

[20] Honari, S., Constantin, V., Rhodin, H., Salzmann, M., Fua, P. (2023). Temporal representation learning on monocular videos for 3D human pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5): 6415-6427. https://doi.org/10.1109/TPAMI.2022.3215307

[21] Fu, H., Gao, J., Liu, H. (2023). Human pose estimation and action recognition for fitness movements. Computers & Graphics, 116: 418-426. https://doi.org/10.1016/j.cag.2023.09.008

[22] Uhlrich, S.D., Falisse, A., Kidziński, Ł., et al. (2023). OpenCap: Human movement dynamics from smartphone videos. PLOS Computational Biology, 19(10): e1011462. https://doi.org/10.1371/journal.pcbi.1011462