Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.075080

ARTICLE

Robust Human Pose Estimation and Action Recognition Utilizing Feature Extraction

Sheng Luo[1], Rashid Abbasi[1,*], Hao Wang[2], Jinghua Xu[3], Dongyang Lyu[4], Aaron Zhang[1], Farhan Amin[5,*], Isabel de la Torre[6], Gerardo Mendez Mezquita[7] and Henry Fabian Gongora[7]

[1]College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, China
[2]Wenzhou Shifeng Technology Co., Ltd., Wenzhou Innovation Center, Wenzhou, China
[3]The State Key Laboratory of Fluid Power Transmission and Control, Zhejiang University, Hangzhou, China
[4]China Electronics Digital Innovation, Hengkuan International Building, Beijing, China
[5]School of Computer Science and Engineering, Yeungnam University, Gyeongsan, Republic of Korea
[6]Department of Signal Theory and Communications, University of Valladolid, Valladolid, Spain
[7]Department of Project Management, Universidad Internacional Iberoamericana, Campeche, Mexico

*Corresponding Authors: Rashid Abbasi. Email: rashidd.abbasi@gmail.com; Farhan Amin. Email: farhanamin10@hotmail.com

Received: 24 October 2025; Accepted: 22 January 2026; Published: 30 March 2026

ABSTRACT: Human pose estimation is crucial across diverse applications, from healthcare to human–computer interaction. Integrating inertial measurement units (IMUs) with monocular vision methods holds great potential for leveraging complementary modalities; however, existing approaches are often limited by IMU drift, noise, and underutilization of visual information. To address these limitations, we propose a novel dual-stream feature extraction framework that effectively combines temporal IMU data and single-view image features for improved pose estimation. Short-term dependencies in IMU sequences are captured with convolutional layers, while a Transformer-based architecture models long-range temporal dynamics. To mitigate IMU drift and inter-sensor inconsistencies, a complementary filtering module is introduced alongside a cross-channel interaction mechanism. Features from the IMU and image streams are then fused via a dedicated fusion module and further refined utilizing a high-precision regression head for accurate pose prediction. Experimental results on benchmark datasets demonstrate that our method significantly outperforms existing techniques in terms of estimation, accuracy, and robustness, validating the effectiveness of our dual-stream architecture.

KEYWORDS: Human pose estimation; dual-stream network; inertial measurement units (IMU)

1 Introduction

Motion capture of the human body is essential for applications, such as gaming (movement of character models, somatosensory games), sports (posture correction), medicine (rehabilitation and posture assessment), VR (Virtual Reality)/AR (Augmented Reality), and film production. Presently, there are three methods: vision-based, IMU-based, and hybrid fusion approaches available in the literature [1,2]. The vision-based human pose estimation relies on multiple cameras and deep learning models, which include RGB cameras [3,4] and depth cameras [5], to achieve high accuracy. However, these methods are highly sensitive to occlusion. Thus, require controlled environments and are unsuitable for indoor scenes with severe occlusions. A novel action scoring algorithm was proposed in [6], where the proposed system integrates angular-based movement analysis with keypoint normalization. The proposed system primarily aims to

minimize angular noise, such as motion blur, etc. However, the proposed system is not validated against high-accuracy motion capture solutions, for instance, Optical Motion Capture (OMC). In addition, the proposed system is designed for single-plane movements. However, multi-view or depth-sensing methods and full 3D joint kinematics systems are not suitable, and thus high cost and the real time remains as challenges for this approach [6].

Currently, vision-based approaches require expensive equipment and specific setups, making them unsuitable for consumer use. It is evident from recent literature that IMU-based techniques, unlike vision-based pose estimation methods, mount IMUs on the human body and are independent of the environment, unaffected by environmental occlusion, and can be used in a wider domain. Their low cost also makes them more suitable for consumer users. However, the disadvantage of this method is that a decrease in the number of IMUs increases pose estimation error, and an increase in the number of IMUs limits the movement of the human body. A hybrid approach is conducted by fusing IMU and visual information to combine the advantages of both sides, but these schemes also have several shortcomings: (1) The noise and drift in the data collected by the IMU will affect the accuracy of attitude estimation. (2) Multi-view will also incur more costs and is unsuitable for outdoor use. (3) Extraction and fusion of IMU and visual features increase system complexity. In this paper, we propose a novel approach to human pose estimation by addressing the limitations of existing methods in both vision-based and IMU-based techniques. Unlike traditional vision-based methods, which suffer from occlusion and require expensive infrastructure, or IMU-based methods, which are constrained by limited sensor numbers and accuracy. This work optimizes the fusion of IMU and visual data to overcome these challenges. The key innovation lies in the feature extraction optimization of IMU data using a three-layer convolutional network, which efficiently aggregates the time information. Thus, it helps in improving pose accuracy in real and dynamic environments. Additionally, the integration of a complementary filter module and a cross-channel interaction module enables the separation and fusion of accelerometer, gyroscope, and gravimeter data, addressing data drift and noise issues. The use of dual-stream data fusion and iterative regression also helps to predict human poses and enhances robustness and precision in the estimation process. Briefly, our framework offers significant improvements over traditional methods by combining the strengths of both IMU and vision data, enabling accurate and real-time human pose estimation in diverse, dynamic environments. The scientific impact of this research is substantial, offering a more affordable, scalable, and accurate solution for applications in gaming, sports, medicine, and VR/AR. In summary, to solve the above problems, the important contribution of the proposed technique is as follows.

- The feature extraction optimization of the IMU was conducted, and the IMU data is modelled by utilizing three convolutional layers to aggregate the time information in a short period of time.
- We have added a complementary filter module and cross-channel interaction module.
- Separating the data of the accelerometer and gravimeter in the IMU, and the gravimeter and gyroscope data are filtered and fused with complementary filters. The relationship between different channels is captured via a cross-channel interaction module.
- Fuse dual-stream data and iterative regression to obtain the predicted pose. Extracted human action features from IMU data and image data, and these features are efficiently fused through the feature fusion module.

The rest of the manuscript is presented as follows. Section 2 reviews human pose estimation and IMU feature extraction. Section 3 describes a comprehensive overview of the proposed scheme. Section 4 reports the quantitative and qualitative evaluation outcomes of our model on the test dataset. Finally, Section 5 discusses the limitations of the proposed approach and outlines possible future directions.

2 Related Work

Vision-based human pose estimation is currently the most widely used method, capturing human movement through cameras and estimating human pose with deep learning [7]. For example, rich temporal features between video frames are exploited to enhance key point recognition, while the spatiotemporal context of key points is encoded to provide sufficient search space, which is then processed by the pose correction network to effectively refine the pose estimation [8,9]. A simple and effective semi-supervised training strategy has been proposed to leverage unlabeled video data for rapid 3D pose prediction in movies, utilizing a fully convolutional model based on extended temporal convolution on 2D key points [10,11]. By proposing a unique encoder–decoder structure with a double-branch decoder, the variable uncertainty of 2D joint positions is solved, and a new large-scale photorealistic synthesis dataset is provided. In addition, a rapid, unified, end-to-end model for estimating 3D human posture, which combines best practices at each stage, has been applied to the evaluation and scoring of rhythmic gymnastics, training, and dance [12–15]. By fusing the camera with an inertial sensor, a higher degree of accuracy in attitude estimation can be achieved; for instance,2D pose detection in each image is correlated with the corresponding IMU equipped by each person, and optimized via a continuous framework to refine a statistical body model pose [16,17]. Furthermore, 3D human pose estimation utilizing multi-view images combined with a small number of IMUs attached to human limbs has been proposed [18,19]. This method first detects a 2D pose from two signals and then lifts it into 3D space [20]. Learn pose embedding from visual occupancy by utilizing multi-channel 3D convolutional neural networks, and semantic 2D pose estimation from Multi-Viewpoint Video (MVV) in discrete volumetric probabilistic visual shells [21]. Presently, human pose estimation based solely on inertial sensors is mainly improved through attitude estimation optimization methods and network models. The pose estimation optimization method is to optimize the process of pose estimation, and there has been some progress in the research of human pose estimation based on IMU at home and abroad. Early studies employed 17 inertial sensors to capture the rotation data of each joint during human movement, to make a more accurate estimation of the human body's posture; however, many sensors will limit the movement of the human body, complicates set up, and increases cost. Following the development of a reliable kinematic model for human studies, a landmark study demonstrated pose estimation utilizing six IMUs for an iterative optimization-based approach [18], but it had to be operated offline, which made real-time applications infeasible. Subsequent work utilized a bidirectional recurrent neural network with six IMUs to directly learn mappings between IMU measurements and human joint rotation, achieving improved accuracy, yet leaving room for further optimization. Recently, a method based on bidirectional recurrent neural networks has been proposed [22]. By dividing the pose estimation into multiple stages, that is, the position of the leaf joint and the position and rotation of the whole joint, estimated step by step from the IMU measurement information, a pose estimation method with higher accuracy is achieved. The overall translation of the wearer is estimated by this step-by-step estimation method. In [23], the authors propose a regular splitting graph network-based approach for 3D human pose estimation. The core idea of the paper was to capture the long-range dependencies between the joints in the body using multi hop neighborhood. They try to learn different modulation vectors for different body joints, as well as a modulation matrix added to the adjacency matrix associated with the skeleton.

3 Proposed Methodology

The overall structure of the network is illustrated in Fig. 1. The network is divided into four main parts: (1) IMU action feature extraction module. The IMU data is input into the IMU fusion module for complementary filtering and fusion after preprocessing operations such as calibration and standardization, and then the IMU action features are extracted through the Transformer module, and the extracted features

are input into the feature fusion module. (2) Image action extraction module. The image data is extracted through the convolutional neural network, and the extracted features are input into the feature fusion module. (3) Feature fusion module that integrates IMU and image action features, and the fused data is input into the regression module. (4) Regression module. The posture that reconstructs human posture from the fused features.

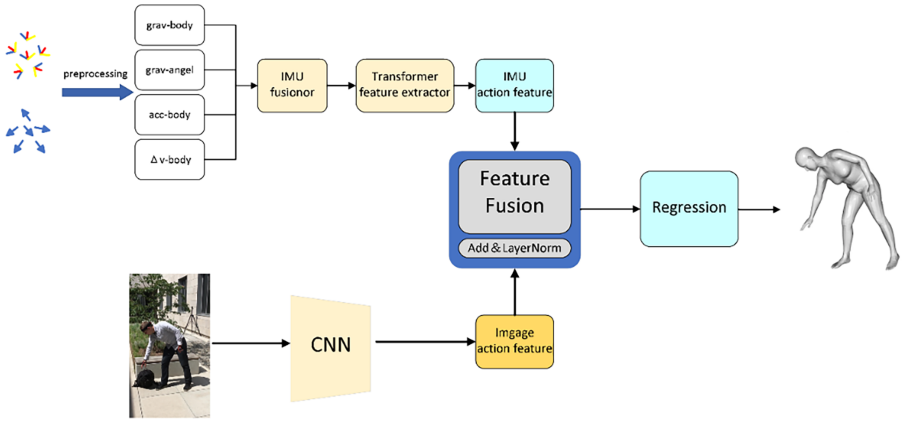

Figure 1: Human pose estimation framework.

3.1 Dual-Stream Feature Extraction

This section describes the feature extraction procedures for IMU and image data. Owing to the difference between IMU and image data, different methods are employed to extract the features and process the IMU data.

3.1.1 IMU Data Feature Extraction

To call the joint point of the waist the root joint point, the joint point of the other five positions is called the leaf joint point. Fig. 1 illustrates the overall flowchart of the IMU feature extraction network, and Fig. 2 presents the detailed network structure of IMU-based feature extraction. IMUs are worn at six locations on the human body: the waist, head, left arm, right arm, left calf, and right calf. The original IMU data are transformed into their independent coordinate system so that the data are relative to the root node in the unified coordinate system. Subsequently, the gravitational acceleration is separated from the IMU acceleration, and the gravitational attitude angle is calculated. The acceleration over a given time interval is then integrated to obtain the body's velocity change, and the above data are then input into the IMU fusion module for further fusion.

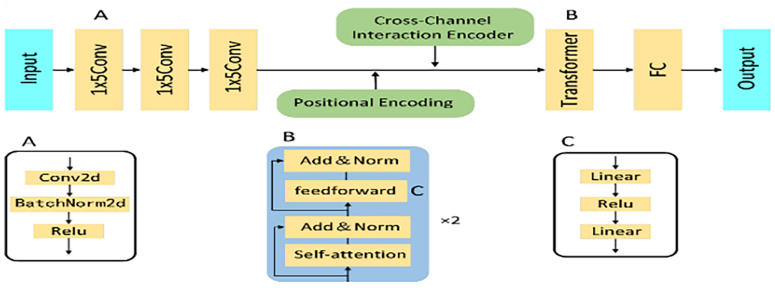

Figure 2: Feature extraction network.

The fused IMU information is subsequently input into the Transformer network to obtain the features of the IMU. At the same time, for the characteristics of IMU data, this section has conducted relevant processing on IMU data to better extract the ground pose features in IMU data. In the IMU-based human pose estimation task, the IMU is utilized to measure partial pose information of the human body. However, gravimeter data is highly susceptible to high-frequency noise, and gyroscope data is primarily affected by low-frequency noise and exhibits significant drift during the integration process. These issues make accurate posture estimation challenging. To address this problem, complementary filtering is applied to the IMU data, where gravimeter data (referring to gravitational acceleration) is processed through a low-pass filter, and gyroscope data is processed through a high-pass filter. The outputs of these two filters are then combined to obtain the pose data as illustrated in Fig. 3.

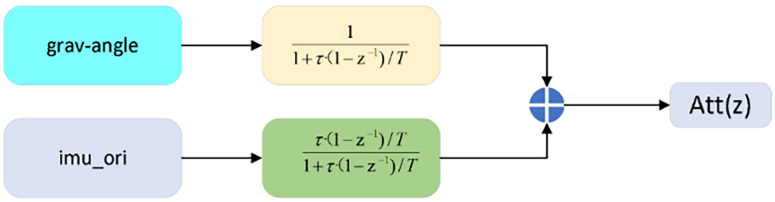

Figure 3: Complementary filtering.

Table 1 is added to clarify the meanings of key symbols in the equation. In this table, the writing of symbols is unified throughout the manuscript. For instance, the term "att" is exclusively used for attitude angle-related variables. The "$attn$" is used for attention weights. The vector and matrix symbols used in formulas, for instance; $(a\_t\hat{}(d, d\hat{}\prime))$.

Table 1: Symbol explanation table.

Symbol	Definition	Application Scenario
z	z-transform operator	Frequency domain analysis of complementary filtering
$att(z)$	z-transform form of the estimated attitude angle	Eq. (1) (Frequency-domain formula of complementary filtering)
$att(k)$	Estimated attitude angle at the k-th timestamp	Eqs. (2) and (3) (Discrete-time formulas of complementary filtering)
$attn_{\mathrm{tk}}$	Attention weight of the k-th channel at the t-th time step	Eq. (5) (Attention mechanism in the IMU fusion module)
$grav\_angle(z)$	z-transform form of the gravitational attitude angle	Eq. (1)
$grav\_angle(k)$	Gravitational attitude angle at the k-th timestamp	Eqs. (2) and (3)
$imu\_ori(z)$	z-transform form of the original IMU attitude angle	Eq. (1)
$imu\_ori(k)$	Original IMU attitude angle at the k-th timestamp	Eqs. (2) and (3)

Fig. 3 shows the discrete frequency formula for the complementary filter, expressed in Eq. (1).

$$\mathrm{att}(z) = \mathrm{grav\_angle}(z) \cdot \frac{1}{1+\tau\cdot(1-z^{-1})/T} + \mathrm{imu\_ori}(z) \cdot \frac{\tau\cdot(1-z^{-1})/T}{1+\tau\cdot(1-z^{-1})/T} \tag{1}$$

where $\mathrm{att}(z)$, $\mathrm{grav}_{\mathrm{angle}(z)}$ and $\mathrm{imu\_ori}(z)$ are the z-transformation form of the estimated gyroscope angular velocity, the gravimeter pose angle, and the pose angle, respectively. τ is a predefined constant, the reciprocal of which is the cutoff frequency, and T is the sampling interval. In addition, by inverse z-transform, the discrete-time formula of the equation can be obtained, as expressed in Eq. (1).

$$\mathrm{att}(k) = \alpha\cdot(\mathrm{att}(k-1) + \mathrm{imu\_ori}(k)\cdot T) + (1-\alpha)\cdot\mathrm{grav\_angle}(k) \tag{2}$$

where k is the timestamp and α denote $\tau/(\tau+T)$. As a recursive formula, Eq. (2) can be further expanded as follows.

$$\begin{aligned} att(k) = {} & \alpha^{k}\cdot grav\_angle(0) + \alpha^{k-1}\cdot[(1-\alpha)\,grav\_angle(1) + \alpha\cdot imu\_ori(1)\,T] + \alpha^{k-2}\cdot \\ & [(1-\alpha)\,grav\_angle(2) + \alpha\cdot imu\_ori(2)\cdot T] + [(1-\alpha)\,grav\_angle(k) + \alpha\cdot imu\_ori(k)\,T] \end{aligned} \tag{3}$$

According to Eq. (3), the coefficient α^{k-i} decreases over time, and the coefficient of gravity attitude angle $(1-\alpha)$ and the rotation coefficient measured by the gyroscope are α added to 1. Based on the above conclusions, the IMU information fusion module can be illustrated in Fig. 3. According to Eq. (3), the complementary filter utilizes only the previous and current information, while the ordinary convolution aggregates the future information, which does not conform to the process of complementary filtering, so the causal convolution module (CConv) is utilized (how is the total convolution). Unlike the normal convolutional kernel, the convolution kernel of causal convolution aggregates weighted data to the final position of the kernel, rather than the central position. As a result, the use of future information can be avoided. Owing to the presence of $grav\_angle(0)$ in Eq. (3), the size of the kernel of the causal convolution of the gravitational attitude angle should be one point larger than the size of the IMU rotation information, and because the coefficients of gravity, attitude angle, and IMU rotation data for each timestamp in Eq. (3) add up to 1, an attention mechanism is added after the causal convolution module. The input to the attention mechanism is the features extracted from the different sensors v_{tk} which can be formalized as follows:

$$\mu_{tk} = \tanh(w_1\cdot v_{tk} + b_1) \tag{4}$$

$$attn_{tk} = \frac{exp(\mu_{tk})}{\sum_k exp(\mu_{tk})} \tag{5}$$

$$f_{tk} = attn_{tk}\cdot v_{tk} \tag{6}$$

Among them, w_1 and b_1 are the parameters of a layer of a multi-layer perceptron (MLP). The normalized weights $attn_{tk}$ are obtained by the softmax function, which is then utilized for modal and scale weighting for each timestamp. (It's not clear where this paragraph is in the process).

3.1.2 Cross-Channel Interaction Module

Accurately achieving delicate human movements with wearable devices typically requires several somatosensory devices that capture activity data through multiple channels. Because the six IMUs are worn on various parts of the human body, these IMUs provide movement information of various parts of the human body, and there is a certain correlation between the information collected by these IMUs. The

interaction among IMU channels in various parts of the human body is relatively implicit and cannot be effectively captured by convolution or linear layers. To address this limitation, a cross-channel interactive encoder is introduced to the channel for capture, which utilizes a self-attention mechanism to effectively reveal the potential channel interaction. Specifically, at each time step t, the normalized correlations between all pairs of sensor channel features are calculated.

$$a_t^{d,d'} = \frac{exp(f(x_t^d)^T g(x_t^{d'}))}{\sum_{d'=1}^{D} exp(f(x_t^d)^T g(x_t^{d'})} \tag{7}$$

where x_t^d and $x_t^{d'}$, f and g, D, and $a_t^{d,d'}$ represent the characteristics of different channels, learned linear transformations, the number of channels, and the channel attention matrix, respectively. Then, on each timestamp, $a_t^{d,d'}$ is multiplied by all channel features.

$$o_t^d = v(\sum_{d'=1}^{D} a_t^{d,d'} h(x_t^{d'})) \tag{8}$$

where o_t^d, and h and v represent the self-attention feature on each timestamp and the learned linear transformations, respectively. Finally, o_t^d is added back to the initial feature via a residual connection. As such, the model has the flexibility to decide whether to utilize the relevant features.

$$r_t^d = o_t^d + x_t^d \tag{9}$$

During training, the attitude estimation model utilizes the cross-channel interaction module to capture the correlations between different sensor channels. These correlations are encoded in self-attention weights and are leveraged at inference time to help support the model's predictions. The Transformer network structure avoids gradient vanishing and explosion, handles long sequences effectively, and outperforms the conventional Recurrent Neural Network (RNN) structure in the extraction of IMU data. The network structure of IMU feature extraction is presented in Fig. 1. As illustrated in Fig. 1, the network of the human pose estimation transformer structure is different from sequential tasks such as natural language processing, and the IMU data has limited action information in the time series. Therefore, this paper utilizes three convolution modules with a size of 1×5 to aggregate the information on the time series, and the aggregated data has richer action information, which is helpful to improve the accuracy of human pose estimation. After the Convolutional Neural Network (CNN) aggregates the time information, the network encodes the position information by adding the position coding, and simultaneously, the potential relationship between the IMUs is obtained by adding the cross-channel interaction module. After that, two identical modules with a self-attention layer and a convolutional layer are passed. The feedforward layer comprises a multi-layer perceptron and a normalization layer that act independently on each timestamp, and because this paper is based on a pre-trained architecture on synthesized data and fine-tuned on real data, a multi-head self-attention mechanism is not utilized to prevent overfitting the synthesized pre-trained dataset.

3.2 Image Data Feature Extraction

The feature extraction of image data was conducted by utilizing the ResNet-18 convolutional neural network. Fig. 2 illustrates the joint structure of the ResNet network. ResNet enables residual connections between any given layer and its subsequent output by allowing the utilization of gradient alternate paths. The residual learning method utilized in this network structure can make the network deeper and enhance easier optimization and accuracy improvement. The image containing the human body is cropped to the size of

224 × 224, and the action features are extracted through the ResNet convolutional neural network, and the extracted features and IMU features are input into the feature fusion module.

3.3 Feature Fusion Module

After extracting the IMU and image features via Transformer and ResNet convolutional neural networks, the features are fused in the feature fusion module. The structure of the feature fusion module is illustrated in Fig. 2. IMU and image-extracted features are first added and input into a feature attention module. The feature attention module is composed of two fully connected layers, two ReLU layers, and a Softmax layer, and the final output is the attention weight of the two features. The output attention weight is multiplied by the added features to the attention feature, and then the maximum feature is obtained by passing through a maximum pooling layer with the IMU feature. Finally, the maximum and image features are added by a residual connection, and the layer normalization operation is conducted to obtain the final fusion feature output. The specific calculation formula is as follows:

$$f_{attention}(t) = \text{softmax}(f_{imu}(t) + f_{img}(t)) \cdot (f_{imu}(t) + f_{img}(t)) \tag{10}$$

$$f_{max}(t) = \text{Maxpooling}(f_{attention}(t), f_{imu}(t)) \tag{11}$$

$$f_{fusion}(t) = \text{LayerNorm}(f_{max}(t), f_{img}(t)) \tag{12}$$

where $f_{imu}(t)$, $f_{img}(t)$, and $f_{attention}(t)$ represent the feature extracted from the IMU at each timestamp, the feature extracted from the image at each timestamp, and the attention feature obtained by adding the feature attention and the two types of features, respectively. $f_{max}(t)$ is the feature after maximal pooling.

3.4 Iterative Regression Module

After feature fusion, the human body's posture features need to be output as the rotation of 15 joints through the regression module. The iterative regression method is employed in the regression module of human pose, and the closest to the real value of the human pose is output through multiple iterations.

As illustrated in Fig. 4, the iterative regression module obtains the iterative pose by adding the fusion feature and the initial predicted pose, initialized with the average pose. The iterative pose is obtained by adding the initial predicted pose through the Dropout layer and two fully connected layers. The initial predicted pose is then updated to the predicted pose, and the above process is repeated. This process is repeated until the predicted pose data is closer to the real pose. The final predicted attitude data is then fed into the Skinned Multi-Person Linear (SMPL) model.

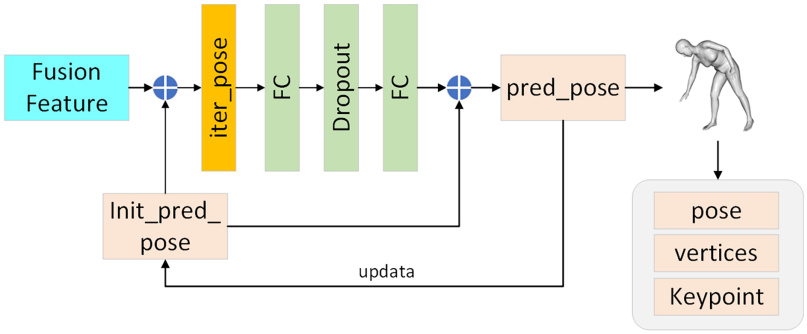

Figure 4: Iterative regression module.

4 Experimental Results and Analysis

4.1 Training Details

The experiment was conducted on a 12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU and RTX 3080 graphics card, utilizing the deep learning framework PyTorch as a network building tool, and the software was trained on PyTorch 1.10.0 and CUDA 11.3. The model utilizes Adam W [5] as the optimizer, and the learning rate at the time of training is the cosine annealing learning rate with a starting learning rate of 0.001. The batch size at the time of training is set to 32, and each experiment is trained for 100 rounds. We set the weight τ of the loss item to 0.01, the φ to 1, and w to 0.01. The human pose estimation by IMU and single-image data, and the experimental data required are image data of human movement and rotation data of each joint point. Designed for 3D pose estimation of markerless multi-camera shots, the TotalCapture dataset [15] is the first dataset with fully synchronized multi-view video, IMU, and Vicon labels for many frames (~1.9 m) of many participants, activities, and viewpoints. The dataset contains many participants who perform different actions and perspectives. It was filmed indoors with eight calibrated 60 Hz HD cameras with a volume of approximately 8 m × 4 m, with a total of four male and one female participant performing four different performances. The IMU part comprises IMU measurements, postural parameters (including acceleration and orientation data), and global translation of approximately 50 min of motion performed by five participants wearing 13 IMUs. The 3 DPW dataset is an outdoor dataset with accurate 3D poses, which can be utilized to solve pose estimation problems. It is the first dataset to contain video footage taken by a mobile phone camera. These include a 60 Hz video sequence of human movements shot with a mobile phone, as well as the corresponding SMPL model parameters. Through the SMPL model parameters and the SMPL model, the corresponding IMU data can be obtained by placing a virtual IMU.

4.2 Test Data and Evaluation Indicators

The model was evaluated utilizing the movements of five participants, walking, performance, and free style in the Total Capture dataset [15]. Metrics include: Sparse Inertial Poser (SIP) Error, the average rotation error of the upper arm and thigh in the global coordinate system, in degrees. Angular error (Ang Err), the global average rotation error of the whole-body joint point, refers to the difference between the angle measured in the pose estimation and the real angle. Position Error (Pos Err) is the average Euclidean distance error of all estimated joints aligned with the root joint (spine), which is the difference between the estimated position and the true position in the positioning system. Mesh Err, the average Euclidean distance error of all vertices of the estimated body mesh aligned with the root joint (spine), which refers to the difference between the mesh model generated in 3D reconstruction or image processing and the real scene. Jitter Err, the average jitter of all body joints in the predicted movement, refers to the instability or fluctuation of the measured value in the time series data.

4.3 Comparative Experiments

In the experimental part of this subsection, the five-evaluation metrics from Section 3.2 are utilized to assess the proposed method. First, this section compares the IMU-based human pose estimation methods on the Total Capture dataset, and the experimental results are presented in Table 2.

Table 2: Comparison with IMU-based methods.

Method	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jit Err (10^2 m/s^3)
SIP	18.54	14.84	7.65	8.6	8.27

(Continued)

Table 2 (continued)

Method	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jit Err (10^2 m/s^3)
DIP	18.79	17.77	9.61	11.34	28.86
TransPose	14.95	12.26	5.57	6.36	1.57
Proposed	11.47	10.45	4.98	5.79	2.63

According to Table 2, the proposed method reduces the SIP error by 3.23 compared to TransPose. Other errors were reduced to a certain extent, except for a slight increase in jitter error. The improvement is attributed to the inclusion of image data with rich features, which compensates for the limited constraints of the sparse IMU. Similarly, the addition of several modules also makes the jitter error increase to a slight extent, and the increased value of the jitter error is within an acceptable range, which will not have a significant impact on the attitude estimation. The method presented in this section is compared with other fusion-based methods, and the comparison is performed on the Total Capture dataset. Test data include walking 2, performance 3, and free style 3 from trained participants 1–3 and untrained participants 4–5. Results are presented in Table 3.

Table 3: Comparison with IMU and image fusion-based methods.

Method	View	IMU	Participant (S1, 2, 3)			Participant (S4, 5)			Average (%)
			W2	A3	FS3	W2	A3	FS3	
Trumble	8	13	30	49	90.6	36	109.2	112.1	70
Malleson	8	13	_	_	65.3	_	64.0	67.0	_
Von Marcard	1	6	_	_	_	_	_	_	39.6
Gilbert	8	13	19.2	42.3	48.8	24.7	58.8	61.8	42.6
LSTM-AE	8	_	13.0	23.0	47.0	21.8	40.9	68.5	34.1
Proposed	1	6	20.6	35.1	59.7	21.6	47.3	60.4	40.8

According to Table 3, the research on the fusion of sparse IMUs with single-view cameras is limited. In this section, we utilize the training and test dataset partitioning scheme introduced by Trumble on the Total Capture dataset for experiments, utilizing the average joint position error in millimeters as an evaluation metric. Annotating the number of views and IMUs clarifies the differences between methods. As presented in Table 3, the method presented in this chapter is superior to the learning-based approach introduced by Trumble et al. [15] utilizes all eight cameras and fuses IMU data with a probabilistic visual shell. It is superior to the others. However, this method is 1.2 mm worse in performance than von Marcard et al. [18]. The main reason is that von Marcard et al.'s [18] approach is video-based on modeling temporal information, with additional correlations between 2D and 3D poses. The results indicate that the proposed method improves human pose estimation accuracy utilizing only single images, without relying on video temporal information. In contrast, LSTM-AE achieves high accuracy through video modelling time information without utilizing IMU, confirming that the time information and multi-view of video are of great help for human posture estimation. To evaluate the effectiveness of fusing IMUs and images in improving the accuracy of 3D pose estimation, an ablation study was conducted on the Total Capture dataset with ground truth annotation.

This study was performed by providing three different inputs, called "IMU only", "image only", and "IMU + image", with the results presented in Table 4.

Table 4: Ablation experiment.

Method	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	JitErr (10^2 m/s^3)
IMU	13.25	11.80	5.69	6.61	1.06
Image	12.30	11.10	5.29	6.15	2.18
IMU and Image	11.47	10.45	4.98	5.79	2.63

Based on the results in Table 4, it can be concluded that the error of the method utilizing the fusion of the IMU and image features is lower than that of the IMU and the image data alone. However, the image-only approach is slightly better than the IMU-only approach. Based on the above experimental results, two conclusions can be drawn. The IMU pose and single-image pose feature can complement each other. The single-image feature can make up for the constraints lacking in the sparse IMU, while the IMU data can make up for the lack of spatial data in a single image, yielding higher overall accuracy. Table 5 shows the r average error in joint position for different body parts. The sparse IMU features, image features have more abundant attitude information, and only IMU data has better stability. Therefore, the estimation error utilizing only image features is smaller than that of utilizing IMU alone, and the jitter error utilizing IMU data alone is the smallest among the three methods.

Table 5: Average error in joint position for different body parts.

Method	Knee	Ankle	Shoulder	Elbow	Wrist
IMU	10.39	10.39	15.96	16.23	16.23
Image	9.95	9.95	14.80	15.34	15.34
IMU and Image	9.47	9.47	13.81	14.44	14.44

Tables 5 and 6 present the average joint rotation error and mean joint position error for each body part in the Total Capture dataset. The average joint selection error and average joint position error of the left and right knees, ankles, shoulders, elbows, and wrists were added together to obtain the average value. Results reveal that the fusion of IMU and image features in the above parts reaches the minimum, indicating that the method proposed in this chapter can achieve the most accurate joint rotation estimation. According to Table 6, the error obtained by combining IMU and image can be minimized in all body parts, and the method is the most accurate in estimating the shoulder position. In the other two separate methods, the image data provides more accurate joint position estimates because it is more informative than the IMU data. This subsection also utilizes the error distribution to obtain the distribution interval of the error of the attitude estimation to make a better judgment on the performance of the model. To ensure the clarity of the pictures, the pictures in this chapter are drawn in the form of dot plots.

Table 6: Average error in joint position for different body parts.

Method	Knee	Ankle	Shoulder	Elbow	Wrist
IMU	7.38	10.52	5.42	9.33	12.08
Image	6.27	9.13	4.30	7.41	10.03
IMU and Image	4.95	7.40	3.30	5.59	7.52

Figs. 5–8 represent the error distribution comparison of the method proposed here and the IMU-only and image-only methods. The abscissa is the error range of each error, and the ordinate is the proportion of the error of each error interval to the total error. As illustrated in the figure, compared with the method that only utilizes a single image and only IMU, the method proposed achieves a relatively accurate error in some timestamps, and has a high proportion of low error intervals in each error, which confirms the effectiveness of the proposed schemes.

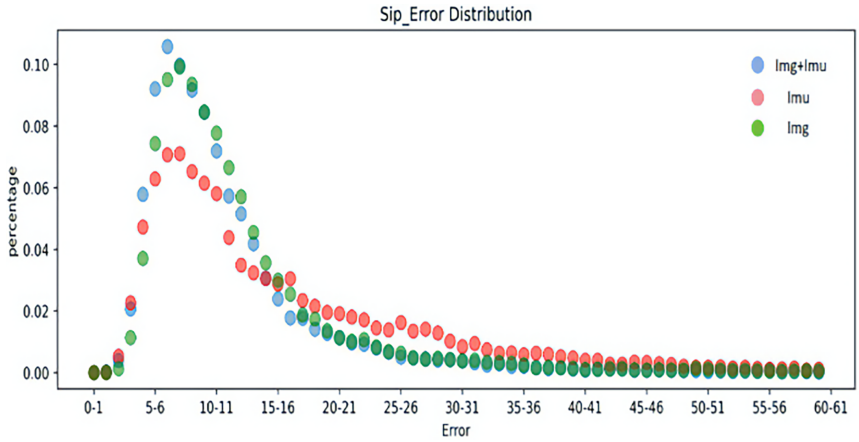

Figure 5: Distribution of SIP errors for different inputs.

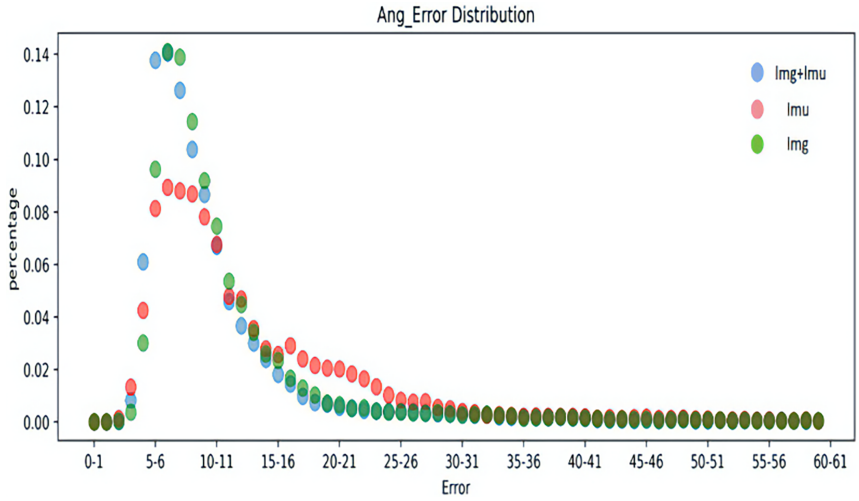

Figure 6: Distribution of joint angle errors for different inputs.

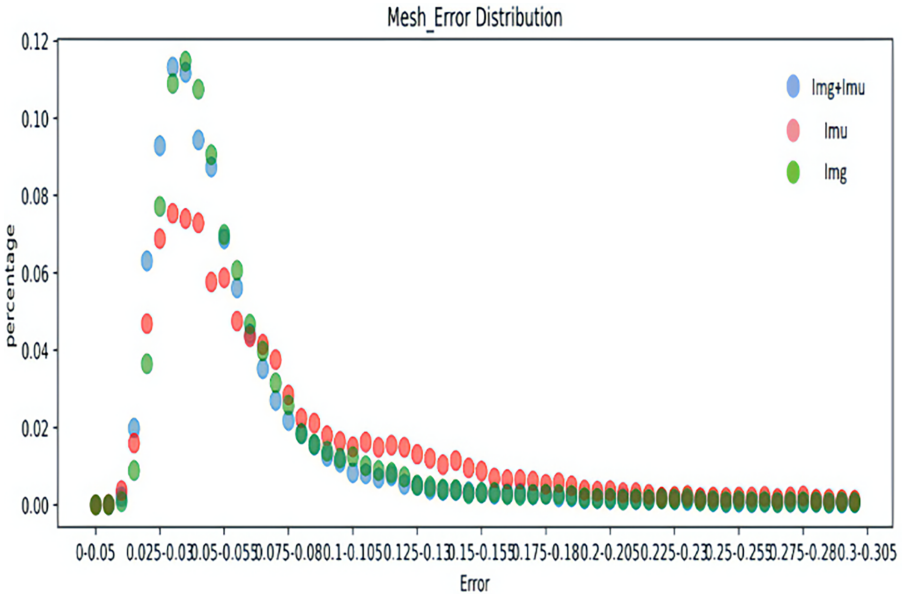

Figure 7: Distribution of joint position errors for different inputs.

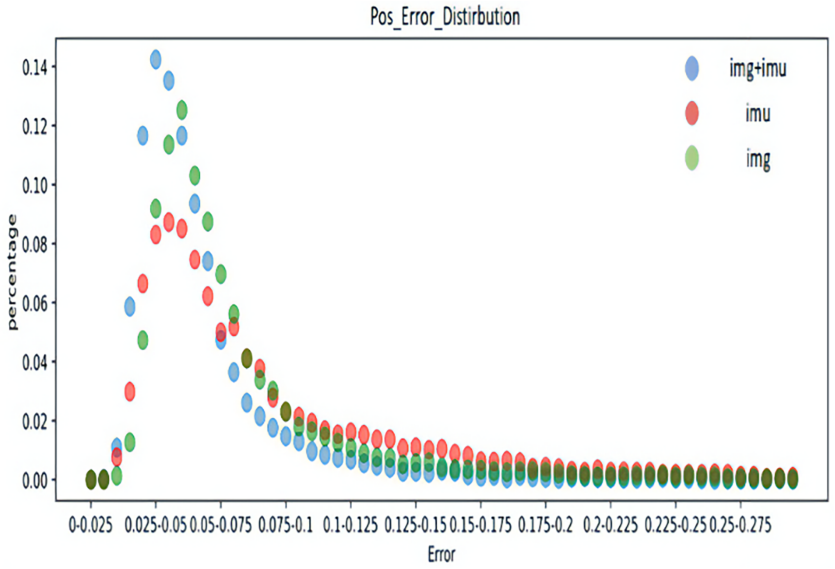

Figure 8: Distribution of mesh vertex errors for different inputs.

4.4 Visual Comparison

To illustrate the improvements of the proposed model on human pose estimation, we first visualize the errors of each technique over time. Then, the frames exhibiting the largest errors are selected to display the corresponding human poses on the SMPL model.

Figs. 9 and 10 are visualizations of SIP error and joint rotation error. Fig. 11 illustrates one of the SIP errors and joint rotation error on the SMPL model. The gray image is the real image; the blue image is the image for pose estimation utilizing the method proposed in this chapter, and the orange image is the image for pose estimation utilizing only the IMU method. From the above images, it can be concluded that the proposed method in this chapter estimates joint rotation more accurately than the IMU-only method. As

illustrated in Fig. 11, the blue model matches the real posture better than the orange model in estimating the rotation of the elbow joint. The average joint position error and grid error of the two indicators of joint position are also visualized to support these results.

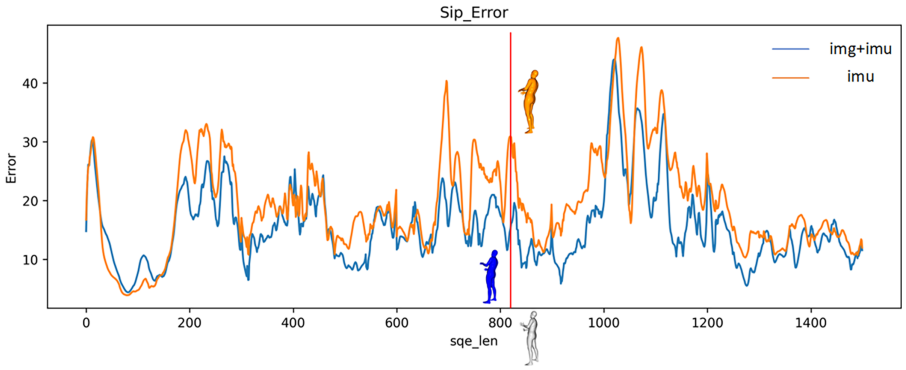

Figure 9: Visualization of SIP error.

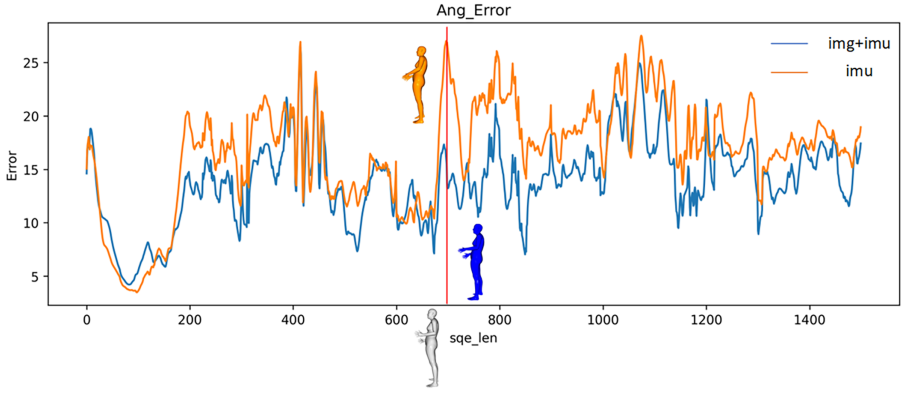

Figure 10: Visualization of joint rotation angle error.

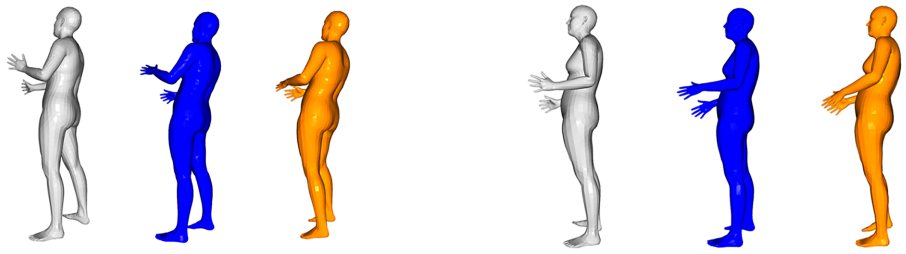

Figure 11: Visualization comparison of SIP error and joint rotation error.

Figs. 12 and 13 are visualizations of the average joint position error and average grid position error. Fig. 14 visualizes one of the frames of the average joint position error and average grid position error on the SMPL model. As illustrated in Figs. 12 and 13, the average joint position error and average grid position of the method presented in this section are lower than those of the IMU-only method in the tested sequence data. Furthermore, the method still maintains a low error rate in several sequences that only utilize the IMU method to produce large estimation errors. Fig. 14 confirms that the proposed method provides notably more

accurate joint position estimates for key body parts, including the legs and head, indicating that it is sufficient to estimate the joint position more accurately.

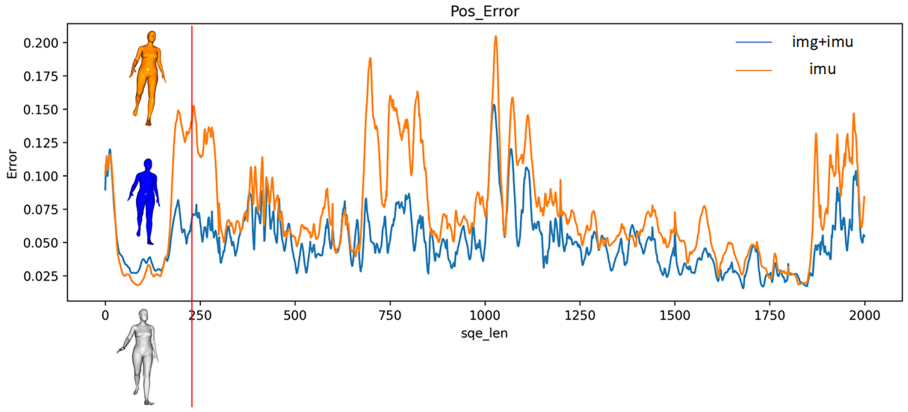

Figure 12: Visualization of average joint position error.

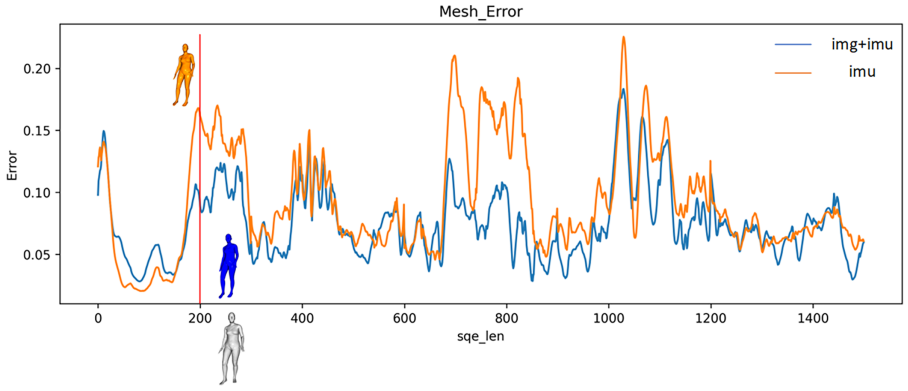

Figure 13: Visualization of average vertex position error.

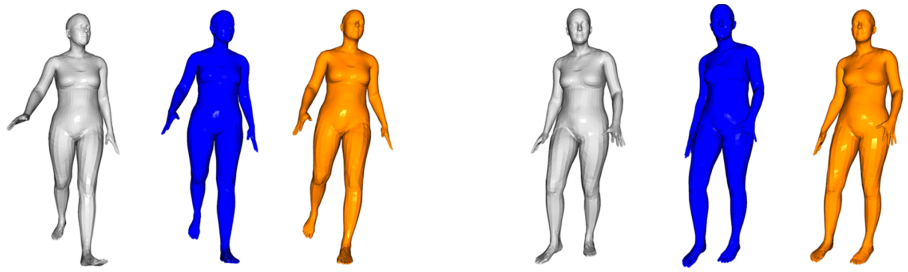

Figure 14: Visualization of average joint position error and average vertex position error.

4.5 Computational Complexity and Resource Requirements

The proposed framework contains multiple modules; the computational overhead of the IMU branch is limited because it operates on low-dimensional inertial sequences. We report computational cost using (i) parameter count, (ii) Floating-Point Operations Per Second (FLOPs), and (iii) practical inference latency and peak GPU memory. All runtime measurements were obtained on our experimental platform (12 vCPU Intel

Xeon Platinum 8255C, NVIDIA RTX 3080, PyTorch 1.10.0, CUDA 11.3) using batch size 1. Let T denote the IMU sequence length, D the number of IMU sensors/channels (here $D = 6$), and C the feature dimension. The complementary filter and causal convolution are linear-time operations with complexity $\mathcal{O}(T \cdot D)$. The cross-channel interaction module computes channel attention per time step with complexity $\mathcal{O}(T \cdot D^2 \cdot C)$; since $D = 6$, this cost is small in practice. The Transformer encoder contributes $\mathcal{O}(T^2 \cdot C)$ in the IMU stream, while the overall end-to-end computation is dominated by the image backbone (ResNet-18 at 224 × 224). A module-wise breakdown is provided in Table 7.

Table 7: Computational complexity and resource requirements.

Module	Params (M)	FLOPs (G)	Latency (ms)	Peak GPU Mem (MB)	Notes
IMU CNN (3 × 1 × 5) + CConv + Complementary Filter	0.135	0.016	0.25	10–20	Linear in T; filter cost negligible
Cross-channel interaction (D = 6)	0.066	0.048	0.35	10–25	O(T·D²·C); small since D = 6
Transformer encoder (2 blocks, single-head)	0.265	0.035	0.55	15–35	Main IMU compute; O(T²·C)
Image branch (ResNet-18, 224 × 224)	11.690	1.800	3.20	160–280	Dominant compute
Fusion block (2-FC attention + pooling + LN)	0.008	0.002	0.10	5–10	Lightweight
Iterative regression (N = 3; 2-FC head)	0.056	0.020	0.20	5–15	Scales linearly with N
Full model	12.220	1.921	4.65	200–350	End-to-end (IMU + image), batch = 1 inference

5 Conclusions

We combined IMU and image data to construct a robust, generalizable model structure. The model integrates IMU data processing, feature extraction, and feature fusion to achieve a reliable human pose estimation. This model enhances the network structure by adding a complementary filtering module and a cross-channel interaction module to optimize the feature extraction of the IMU. Dual-stream features are then fused, and the human pose is predicted via the iterative regression module. However, it may still have the problem of insufficient constraints. Adding a physics module could simulate the physical information constraints of the human body during motion. In the future, more advanced fusion approaches, such as end-to-end multimodal networks, can be explored to better leverage the advantages of both data sources.

Acknowledgement: This research is funded by the European University of Atlantic.

Funding Statement: No additional external funding was received beyond the support provided by the European University of Atlantic.

Author Contributions: Sheng Luo and Farhan Amin: conceptualization and writing—original draft preparation. Rashid Abbasi: methodology. Jinghua Xu: software. Dongyang Lyu, Hao Wang and Gerardo Mendez Mezquita: validation. Henry Fabian Gongora and Aaron Zhang: software and writing—original draft preparation. Isabel de la Torre: software, investigation, and conceptualization. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analyzed during the current study are available in the TotalCapture Dataset, https://cvssp.org/data/totalcapture/data/.

Ethics Approval: This study was conducted using publicly available. Therefore, formal ethics approval was not required, or not applicable.

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

References

1. Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1302–10. doi:10.1109/CVPR.2017.143.
2. Yan L, Du Y. Exploring trends and clusters in human posture recognition research: an analysis using CiteSpace. Sensors. 2025;25(3):632. doi:10.3390/s25030632.
3. Güler RA, Neverova N, Kokkinos I. DensePose: dense human pose estimation in the wild. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7297–306. doi:10.1109/CVPR.2018.00762.
4. Huang Y, Kaufmann M, Aksan E, Black MJ, Hilliges O, Pons-Moll G. Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time. ACM Trans Graph. 2018;37(6):1–15. doi:10.1145/3272127.3275108.
5. Kinga D, Adam JB. A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015); 2015 May 7–9; San Diego, CA, USA.
6. Ullah R, Asghar I, Nawaz R, Stacey C, Akbar S, Bishop P. A real time action scoring system for movement analysis and feedback in physical therapy using human pose estimation. Sci Rep. 2025;15(1):44784. doi:10.1038/s41598-025-29062-7.
7. Liu Z, Chen H, Feng R, Wu S, Ji S, Yang B, et al. Deep dual consecutive network for human pose estimation. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 525–34. doi:10.1109/cvpr46437.2021.00059.
8. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. SMPL: a skinned multi-person linear model. In: Seminal graphics papers: pushing the boundaries. Vol. 2. New York, NY, USA: ACM; 2023. p. 851–66. doi:10.1145/3596711.3596800.
9. Mahmood N, Ghorbani N, Troje NF, Pons-Moll G, Black M. AMASS: archive of motion capture as surface shapes. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 5441–50. doi:10.1109/iccv.2019.00554.
10. Nguyen HC, Nguyen TH, Scherer R, Le VH. Unified end-to-end YOLOv5-HR-TCM framework for automatic 2D/3D human pose estimation for real-time applications. Sensors. 2022;22(14):5419. doi:10.3390/s22145419.
11. Pavllo D, Feichtenhofer C, Grangier D, Auli M. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 7745–54. doi:10.1109/CVPR.2019.00794.
12. Chen H, Fan R. Improved convolutional neural network for precise exercise posture recognition and intelligent health indicator prediction. Sci Rep. 2025;15(1):21309. doi:10.1038/s41598-025-01854-x.
13. Wu Y, Liu J. Research on college gymnastics teaching model based on multimedia image and image texture feature analysis. Discov Internet Things. 2021;1(1):15. doi:10.1007/s43926-021-00015-6.
14. Tome D, Peluse P, Agapito L, Badino H. xR-EgoPose: egocentric 3D human pose from an HMD camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 7728–38.
15. Trumble M, Gilbert A, Malleson C, Hilton A, Collomosse J. Total capture: 3D human pose estimation fusing video and inertial sensors. In: Proceedings of the British Machine Vision Conference; 2017 Sep 4–7; London, UK. doi:10.5244/c.31.14.

16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems 30 (NIPS 2017). Red Hook, NY, USA: Curran Associates, Inc.; 2017. doi:10.65215/ctdc8e75.
17. von Marcard T, Henschel R, Black MJ, Rosenhahn B, Pons-Moll G. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: Proceedings of the Computer Vision—ECCV 2018. Cham, Switzerland: Springer International Publishing; 2018. p. 614–31. doi:10.1007/978-3-030-01249-6_37.
18. von Marcard T, Rosenhahn B, Black MJ, Pons-Moll G. Sparse inertial poser: automatic 3D human pose estimation from sparse IMUs. Comput Graph Forum. 2017;36(2):349–60. doi:10.1111/cgf.13131.
19. Wei X, Zhang P, Chai J. Accurate realtime full-body motion capture using a single depth camera. ACM Trans Graph. 2012;31(6):1–12. doi:10.1145/2366145.2366207.
20. Xia D, Zhu Y, Zhang H. Faster deep inertial pose estimation with six inertial sensors. Sensors. 2022;22(19):7144. doi:10.3390/s22197144.
21. Yi X, Zhou Y, Xu F. TransPose: real-time 3D human translation and pose estimation with six inertial sensors. ACM Trans Graph. 2021;40(4):1–13. doi:10.1145/3450626.3459786.
22. Zhang Z, Wang C, Qin W, Zeng W. Fusing wearable IMUs with multi-view images for human pose estimation: a geometric approach. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 2197–206. doi:10.1109/cvpr42600.2020.00227.
23. Hassan MT, Ben Hamza A. Regular splitting graph network for 3D human pose estimation. IEEE Trans Image Process. 2023;32:4212–22. doi:10.1109/tip.2023.3275914.