Mamba-based 3D Human Pose Estimation Algorithm for Track and Field Training

Sangbing Tsai*

International Engineering and Technology Institute, Hong Kong; klj0418@gmail.com

*Corresponding Author: klj0418@gmail.com

DOI: https://doi.org/10.30211/JIC.202503.014

Submitted: Aug. 18, 2025 Accepted: Oct. 08, 2025

ABSTRACT

In track and field training, the accuracy of athletes' technical movements directly affects their sports performance and sports injury prevention. Traditional posture analysis methods mainly rely on the coach's empirical observation, but this approach suffers from the shortcomings of limited evaluation angles and untimely feedback, which makes it difficult to provide accurate and real-time movement optimization especially in high-speed and high-dynamic sports scenarios. To address this problem, this paper proposes PoseFusionNet, a 3D human posture estimation model that combines OpenPose and Mamba algorithms. The model maps the 2D joint data extracted by OpenPose to 3D space and combines the timing information to improve the posture estimation's accuracy and robustness. Experimental results show that PoseFusionNet can still maintain high accuracy under multiple challenging scenarios (e.g., high-speed motion, occlusion, and illumination changes) and has real-time feedback capability, which can satisfy the dual demands of accuracy and real-time performance in track and field training. The model provides a new technical tool for intelligent sports training, which can help to improve the training efficiency of athletes and reduce sports injuries.

Keywords: 3D pose estimation, OpenPose, Mamba algorithm, Athlete motion analysis, Track and field training, Real-time feedback

1. Introduction

In track and field training, the accuracy of athletes' technical movements directly impacts their performance and career longevity. Whether in sprinting, long jump, or throwing events, optimizing movement posture is crucial for enhancing athletic performance and preventing injuries. This is especially challenging in high-speed sports environments, where athletes' movements are highly complex and dynamic. Traditional training methods largely rely on coaches' visual observation and experience to correct techniques, which suffer from inherent limitations. First, a coach's ability to observe is constrained by viewing angles, subjective experience, and perceptual speed, making it difficult to provide comprehensive and accurate assessments of each athlete's movements in real time. Second, the rapid and continuous changes in posture during high-intensity training are often too

fleeting for the human eye to capture in detail, resulting in a lack of timely and precise feedback for technical improvement.

Therefore, achieving accurate and real-time analysis of athletes' movement postures has become essential for improving the efficiency and effectiveness of track and field training.

In recent years, advances in computer vision, deep learning, and sensor technology have made vision-based human pose estimation an increasingly important research direction in sports training. Deep learning models enable computers to extract key skeletal points from sports videos, allowing for automated analysis and optimization of athlete posture. Such automated systems can significantly reduce the workload of coaches while providing athletes with scientific and personalized training feedback.

Human pose estimation technologies are generally categorized into 2D and 3D approaches. Traditional 2D pose estimation can effectively extract two-dimensional keypoints—such as joint locations—from images, but it fails to capture the body's spatial relationships and posture in three dimensions. As a result, 2D methods are insufficient for comprehensive technical analysis, especially in fast-paced or complex motions where the lack of depth information may lead to inaccurate assessments. In contrast, 3D pose estimation reconstructs the three-dimensional coordinates of human keypoints from two-dimensional images, offering more complete and accurate kinematic information. This is critical for performance analysis, movement optimization, and injury prevention. However, 3D pose estimation still faces challenges in maintaining real-time performance and high accuracy while addressing common issues such as occlusion and motion blur.

Among the various pose estimation algorithms, OpenPose and Mamba have attracted significant attention and are widely used. OpenPose is a 2D pose estimation method based on convolutional neural networks (CNNs), capable of accurately detecting 2D human keypoints in single images or videos while supporting multi-person estimation. It is known for its high speed and accuracy, enabling real-time processing of large-scale data, and has been extensively applied in sports analysis and live monitoring. Nonetheless, since OpenPose only provides 2D output, it must be integrated with 3D pose estimation techniques to obtain spatial joint positions.

Mamba is a 2D to 3D pose estimation algorithm that uses deep neural networks (including graph convolutional networks and time series modeling technology) to map the 2D joint position data extracted by OpenPose to 3D space to generate accurate three-dimensional human pose information. Mamba effectively solves the inaccuracy in single-frame image estimation by combining time series information (such as the athlete's historical action data) and global space information and improves the accuracy and robustness of 3D pose estimation. Combining OpenPose and Mamba can give full play to the advantages of both, make up for the shortcomings of a single algorithm, and achieve efficient and accurate real-time 3D human pose estimation.

Although these two technologies have high theoretical value and application potential, in actual sports training, how to efficiently combine them to solve problems such as occlusion, image blur, and real-time feedback in fast motion is still a challenge to be solved. Existing research focuses on the implementation of theoretical algorithms and applications in a single scenario. For high-dynamic and

high-speed motion scenes in track and field training, there is still a technical gap in how to achieve high-precision and high-real-time 3D posture estimation. Therefore, the goal of this study is to combine OpenPose and Mamba to build a system that can estimate the 3D posture of athletes in track and field training in real time and accurately and verify its effect and application value in actual sports training.

To this end, this paper proposes a new posture estimation system named PoseFusionNet. This system combines the advantages of OpenPose in 2D posture estimation and the accuracy of Mamba in 3D posture recovery. Through innovative network structure design and optimization strategies, it achieves high-precision, low-latency real-time 3D human posture estimation. The innovation of PoseFusionNet lies in how to efficiently fuse 2D and 3D information to achieve accurate 3D pose recovery in complex motion scenes, especially in high-speed motion and dynamic changes, overcoming the occlusion and motion blur problems in traditional methods.

The innovations of this study are:

- Algorithm fusion: Combine the OpenPose algorithm based on 2D pose estimation with the Mamba algorithm based on 3D pose estimation and make full use of the advantages of both to ensure real-time performance and improve estimation accuracy.
- Application scenario: Take track and field training as the application scenario, focus on motion analysis and optimization in running, long jump and other projects, and explore how to use 3D pose estimation to help athletes improve their technical movements.
- Technical challenges: Solve problems such as occlusion and noise in high-dynamic sports environments and ensure the availability and stability of the system in real training environments.

This study will verify the effectiveness of the system through a series of experiments. We will introduce the data collection method, experimental design, evaluation indicators and experimental results, and analyze and discuss the results in detail. Finally, this paper will also explore the application prospects of the system in actual training and propose possible future improvement directions.

Through this study, we hope to provide new technical means for track and field training, help athletes achieve more accurate motion analysis and optimization during training, thereby improving sports performance and reducing sports injuries, and provide new research ideas and practical experience for the application of deep learning-based human posture estimation technology in sports training.

2. Literature Review

2.1 Evolution and Challenges of 2D and 3D Human Pose Estimation Technology

As an important research field in computer vision, human pose estimation initially focused on 2D pose estimation. Early research focused on extracting key points of the human body (such as 2D coordinates of joints and body parts) from static images and analyzing the human body's motion posture through the relative positions of these joints[1, 2]. Classical methods of 2D pose estimation

technology, such as deep convolutional neural networks (CNNs) and traditional image-based algorithms, have achieved remarkable results in a variety of application scenarios[3]. For example, CNN-based 2D pose estimation algorithms such as OpenPose can accurately extract the joint positions of the human body in a single image or video[4, 5], which provides an effective tool for real-time monitoring and analysis of athletes' technical movements[6].

However, although 2D pose estimation has greatly improved in accuracy and efficiency, it still has certain limitations[7]. 2D pose estimation cannot capture the depth information of the human body in three-dimensional space, especially when it comes to complex three-dimensional movements of athletes, 2D algorithms seem to be unable to cope with it[8]. For example, the relative position changes of joints, the rotation and tilt of body postures[9], relying solely on two-dimensional coordinates often cannot accurately reflect their true three-dimensional spatial relationships[10]. Therefore, researchers have begun to turn their attention to 3D pose estimation technology, aiming to reconstruct a more accurate pose model by restoring the 3D structure of the human body from 2D images[11].

3D pose estimation can restore the 3D joint position of the human body from images or videos. Compared with traditional 2D methods, it can provide more accurate motion analysis, especially when dealing with fast dynamic or complex movements[11]. In recent years, 3D pose estimation algorithms based on deep learning have been widely used, among which convolutional neural networks (CNN) and graph convolutional networks (GCN) have become mainstream technical means. By combining multi-view data, depth images, time series information, etc.[12], many studies have achieved efficient estimation of 3D human pose. Despite this, existing 3D pose estimation technology still faces some challenges that need to be solved, especially in dynamic scenes. Problems such as occlusion, motion blur, and joint overlap often lead to a decrease in the accuracy of the estimation results[12, 13]. These problems are particularly prominent in sports training scenarios, especially when athletes are moving at high speed or have complex movements. How to provide high-precision real-time 3D pose feedback is still a technical bottleneck.

2.2 Human Pose Estimation and Time Series Modeling Based on Deep Learning

The introduction of deep learning has had a revolutionary impact on human posture estimation, especially the successful application of convolutional neural networks (CNN) in 2D posture estimation, which has promoted the development of this field[14, 15]. By training deep networks, computers can automatically extract complex spatial features from images to achieve efficient and accurate posture estimation[16]. With the improvement of computing power, deep learning technology has gradually been applied to 3D posture estimation[17]. Many studies use deep convolutional networks to infer the three-dimensional coordinates of the human body from single-view images or videos[18], especially through time series modeling and time series analysis, which can further improve the prediction ability of complex movements[19].

Time series modeling technology, especially long short-term memory networks (LSTM) and time series convolutional networks (TCN), have been widely used in the field of human posture estimation. These technologies introduce the time dimension[19], allowing the algorithm to capture

the changing trend of movement from a series of continuous images or video frames, thereby improving the estimation accuracy of dynamic posture. For example, in the training process of athletes, the coherence and timing of movements are often crucial to the accurate analysis of posture[20, 21]. The use of time series modeling can effectively reduce the noise and errors in single-frame images, especially in the case of occlusion and fast movement. However, although time series modeling can improve the accuracy of dynamic posture estimation to a certain extent, it still faces the problems of data synchronization[22], high computing resource requirements and high training complexity. In addition, time series modeling often relies on a large amount of high-quality training data, while the dynamic scene data in sports training is expensive to obtain and difficult to label, which also limits its application scope.

2.3 Application of Athlete Posture Estimation in Training Optimization

In athlete training, posture estimation technology is mainly used in motion analysis, technical optimization and sports injury prevention. Traditional sports training relies on the visual observation and feedback of coaches[23, 24]. Although coaches are experienced, the observation ability and reaction speed of human eyes are limited and cannot accurately capture every detail. In recent years, automated posture estimation technology can help coaches find problems in athletes' movements in time and provide personalized technical guidance based on data[25]. This technology can be applied to a variety of sports such as track and field, basketball, and football to help athletes accurately adjust their postures, thereby improving sports performance and reducing the risk of injury[26].

Especially in track and field training, athletes' technical movements (such as sprint start, long jump, throwing movements, etc.) are usually highly dynamic[27, 28], with fast and complex rhythm changes. By obtaining the athlete's 3D posture in real time, the coach can analyze his movements in detail and adjust the training content in time. However, in a high-dynamic environment, especially in the case of fast movement and rapid changes in athlete posture[29], posture estimation technology still faces great challenges. Factors such as occlusion, motion blur, and background noise often reduce the accuracy of the algorithm, especially when using traditional visual methods[30], the estimated joint position is easily disturbed. Although some algorithms can deal with these problems by optimizing deep learning models, there is still a lack of efficient solutions in applications of high-intensity exercise and real-time feedback[31].

Therefore, how to improve the accuracy and real-time performance of posture estimation in a dynamic training environment has become one of the current research hotspots. Most of the existing technologies focus on improving the accuracy of athlete motion analysis[32], and how to ensure high accuracy while dealing with the blur and occlusion problems caused by rapid motion is still an important research direction.

2.4 Multimodal Data Fusion and Cutting-Edge Technology for Efficient Posture Estimation

With the advancement of sensor technology, multimodal data fusion has gradually become an effective means to improve the accuracy and robustness of posture estimation. In addition to traditional RGB images, sensors such as depth maps[33], infrared imaging and IMU (inertial

measurement unit) have also begun to be widely used in human posture estimation. Multimodal data fusion can complement information between different data sources and help the system overcome the limitations of a single sensor in complex scenes[34, 35]. For example, depth maps can provide more accurate spatial positions, while infrared images can provide additional visual information in low-light environments. By fusing different types of data, the accuracy and robustness of the estimation results can be effectively improved[36].

In addition, the application of time series modeling technology also plays an important role in multimodal data fusion. By combining multimodal data and time series information, researchers can better capture the changes in athletes' movements and provide more coherent and accurate posture estimation[37, 38]. Although this technology has made significant progress in laboratory environments, in actual training environments, how to deal with the synchronization of multimodal data, how to solve noise interference[39], and how to optimize algorithms to ensure real-time performance are still the main challenges for technical implementation.

Most existing research focuses on improving accuracy and processing speed by optimizing deep learning algorithms[40, 41]. However, in practical applications, how to ensure that these optimization solutions can effectively cope with the rapid changes, environmental interference and computing resource limitations in athletes' high-intensity training still requires more exploration[42]. Therefore, how to further improve the accuracy and stability of multimodal data fusion while ensuring real-time performance is an important direction for future research.

3. Research Design

3.1 Overview of Our Network

At present, although many studies have made important progress in the field of 2D and 3D pose estimation, there are still many challenges. How to accurately convert 2D joint positions into pose information in 3D space, especially in complex dynamic scenes such as athlete training analysis, is still an urgent problem to be solved. Existing methods usually rely on 2D pose estimation results to infer 3D positions, but due to the influence of joint occlusion, motion complexity and data noise, the accuracy and robustness of traditional 3D pose estimation methods in these environments are poor. To solve these problems, this paper proposes a new model, PoseFusionNet, which combines the 2D joint position estimation of OpenPose and the deep learning architecture of the Mamba model, aiming to improve the accuracy and robustness of 3D pose estimation through multimodal information fusion.

The core design objective of PoseFusionNet is to enhance spatial modeling capability during 3D pose reconstruction by leveraging accurate 2D joint position inputs, along with convolutional neural networks (CNNs) and graph convolutional networks (GCNs). The model architecture consists of the following key components: an OpenPose module that extracts 2D joint positions from input images; a CNN that further captures spatial features and learns relative positional relationships between joints; and a GCN module that strengthens spatial dependencies among joints and models complex structural relationships. A regression network subsequently maps these deep features into 3D space, precisely recovering the 3D coordinates of each joint.

Within the network pipeline, each module fulfills a critical function. The process begins with the OpenPose module processing the 2D input image to generate accurate 2D joint coordinates. OpenPose's high precision in 2D pose estimation ensures reliable initial joint localization, establishing a solid foundation for 3D reconstruction. The extracted 2D joint data is then passed through CNN, where multi-layer convolutional operations learn discriminative spatial features—such as inter-joint motion patterns and structural relationships, thereby providing rich semantic information for 3D recovery. Subsequently, GCN refines the modeling of joint dependencies, capturing both global and local spatial structures. Graph convolution operations enable the model to better interpret spatial configurations between joints, significantly improving robustness when reconstructing complex motions and postures. The final regression network outputs the 3D position of each joint, achieving accurate three-dimensional pose reconstruction.

During the training process, the PoseFusionNet model adopts an end-to-end training strategy and uses a large-scale annotated dataset for supervised learning. In terms of training objectives, the model optimizes the weighted combination loss function of the two-dimensional joint reconstruction error and the three-dimensional posture recovery error to ensure that the model can not only accurately restore the three-dimensional posture, but also accurately fit the two-dimensional input. To improve the stability and generalization ability of training, this study introduces data enhancement technology, including operations such as rotation, scaling, and cropping of images to increase the adaptability of the model to different postures and movements. At the same time, the Adam optimizer is used to train the model to ensure that the optimization process in high-dimensional feature space is both efficient and stable.

Compared with traditional single models, PoseFusionNet has several significant advantages. First, with the high-quality 2D joint positions provided by the OpenPose module, the model obtains relatively accurate prior information at the input stage, thereby reducing the error propagation in the 3D pose estimation process. Second, the combination of CNN and GCN enables the model to deeply mine spatial features, especially in the case of complex motion or joint occlusion, which can significantly improve the robustness of 3D pose estimation. In addition, through the design of the regression network, PoseFusionNet can accurately map 2D joint information to 3D space, improving the accuracy of 3D pose recovery.

This study expects PoseFusionNet to play an important role in multiple application scenarios, especially in the field of athlete training and motion analysis. In track and field training, the PoseFusionNet model can analyze the athlete's movements in real time and provide data support and training feedback to the coach through accurate 3D pose recovery. Through this technology, the coach can not only monitor the training effect of the athlete but also provide more targeted action improvement suggestions for the athlete. With the continuous enrichment of training data and application scenarios, PoseFusionNet can also be applied to multiple fields such as virtual reality and robot control, showing broad application prospects

3.2 OpenPose

The core principle of the OpenPose algorithm is to extract features from the input image through

a deep convolutional neural network (CNN) and generate heat maps of key points of the human body through regression methods. The algorithm gradually extracts the spatial features of the image through multiple convolutional layers and locates each joint of the human body through feature maps. During the processing, OpenPose first performs a convolution operation on the input image to extract low-level feature information, then gradually generates a heat map of each joint position and finally restores the complete human skeleton by optimizing the joint connection relationship. Next, a series of formulas will be used to explain this process in detail.

First, the input image I is processed by multiple layers of convolutional neural network (CNN) to obtain the feature map F, where the convolution operation of each layer can be expressed as:

$$F = \text{Conv}(I, W) + b \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{[Formular 1]}$$

where I am the input image, W is the convolution kernel, b is the bias term, Conv represents the convolution operation, and the obtained feature map F contains the spatial information in the image.

Next, OpenPose generates the heat map H_j of the joints through a regression method (each joint corresponds to a heat map), which indicates the possibility of the joint at each pixel position in the image. The generation of the heat map is expressed by the following formula:

$$H_j = \sigma(\text{Conv}(F, W_j) + b_j) \cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{[Formular 2]}$$

where H_j is the heat map of joint j, W_j and b_j are the convolution kernel and bias terms associated with joint j, and σ is an activation function (usually ReLU or Sigmoid) used to enhance the nonlinear representation of features.

To further optimize the recovery of the human skeleton, OpenPose introduces the spatial relationship between joint pairs and improves the pose estimation by connecting these joints. In this process, the connection prediction of the joint pair is expressed by the following formula:

$$P_{ij} = \sum_{p \in P} \alpha_{ij}(H_i, H_j, p) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{[Formular 3]}$$

where P_{ij} is the connection prediction between joints i and j, P is a selected pixel region, α_{ij} is a parameter representing the strength of the joint connection, H_i and H_j are the heat maps corresponding to joints i and j, and p is the pixel position.

To obtain a globally optimized skeleton structure, OpenPose further optimizes the connection between joint pairs to minimize the error of joint connection. This process is carried out through the following optimization formula:

$$S = \arg\min_{S'} \sum_{i,j} \| P_{ij}(S') - P_{ij}(S) \|^2 \cdots\cdots\cdots\cdots \text{[Formular 4]}$$

where S is the final skeleton structure, $P_{ij}(S')$ is the optimized joint pair connection prediction, and $P_{ij}(S)$ is the preliminary estimated joint pair connection prediction. The goal is to obtain a globally optimized skeleton structure by minimizing the prediction error.

Through the above steps, OpenPose can accurately extract the 2D joint position and skeleton structure of the human body from the input image. To provide a clear understanding of how OpenPose fits within the PoseTrackNet model, a schematic diagram of the model structure is shown in Figure 1. This diagram illustrates the flow of data through the various components of the model, highlighting

the integration of pose estimation and tracking processes.

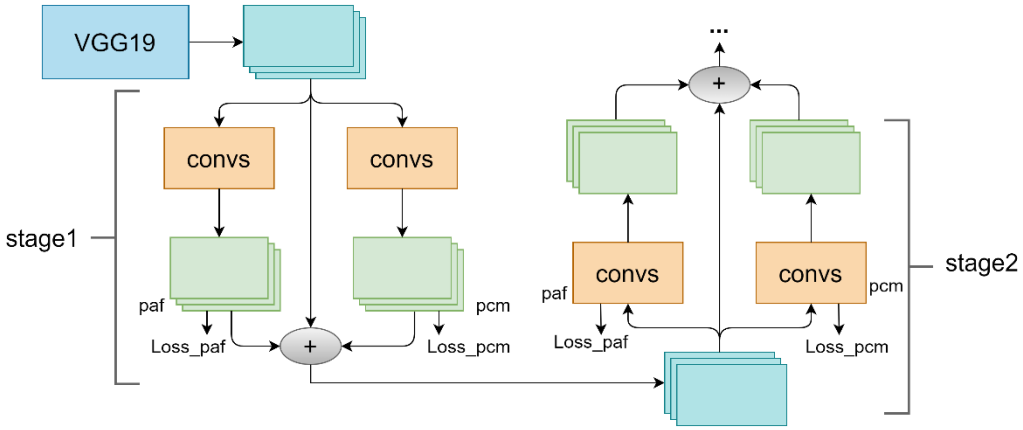

Figure 1. The structure of OpenPose

OpenPose has a wide range of applications in the field of human posture estimation, especially in sports training, human behavior analysis, intelligent monitoring and virtual reality. In sports training, OpenPose can capture and analyze athletes' movements in real time, helping coaches to find problems in athletes' technical movements in a timely manner and provide accurate training feedback. In track and field training, OpenPose extracts the key points of athletes through video streams and analyzes the details of their movements, such as running start, jumping and other technical movements, thereby providing data support for athletes and helping them improve their movement techniques.

In addition, OpenPose's real-time and multi-body detection capabilities make it also excellent in multi-person sports scenes. It can distinguish the key points of different athletes in complex backgrounds and perform efficient posture estimation, which is particularly important for the simultaneous training of multiple athletes. Compared with traditional manual analysis methods, OpenPose can greatly improve analysis efficiency and accuracy, reduce manual intervention, and improve the scientific nature of training.

In this paper, OpenPose, as the core module of 2D human posture estimation, undertakes the task of extracting athlete key points from video frames. These two-dimensional key points will be passed as input data to the subsequent 3D posture estimation module to provide necessary information for 3D posture reconstruction. By using OpenPose for efficient 2D pose estimation, this paper can accurately capture the motion characteristics of athletes in complex sports scenes and further convert them into joint positions in three-dimensional space. This process provides scientific data support for the analysis of athletes' technical movements and provides a basis for subsequent training optimization and injury prevention.

Therefore, OpenPose plays a vital role in this paper. It not only provides accurate 2D key point data but also lays the foundation for the accuracy improvement and real-time guarantee of the entire 3D pose estimation system. By combining with the subsequent Mamba model, OpenPose effectively

combines 2D and 3D pose estimation, promoting the application of human pose estimation technology in sports training

3.3 Mamba

Mamba is a deep learning-based 3D human pose estimation algorithm that aims to infer the human skeleton structure in 3D space from 2D keypoint data. Traditional 3D pose estimation methods usually rely on direct 3D data input or complex multi-view images, while Mamba utilizes 2D keypoint information generated by 2D pose estimation technology (such as OpenPose) and maps it through deep neural networks to derive the corresponding 3D pose. This method not only improves computational efficiency but also demonstrates high accuracy and robustness in a variety of practical application scenarios.

The core architecture of Mamba integrates convolutional neural networks (CNNs) and graph convolutional networks (GCNs) to enhance 3D pose estimation. The CNN component extracts richer feature representations from input 2D keypoints, while the GCN explicitly models the spatial relationships between human joints. By combining these two modules, Mamba effectively captures structural dependencies among joints, thereby improving the accuracy of 3D pose reconstruction.

Traditional 2D pose estimation methods typically obtain joint coordinates from images, yet such 2D information alone is insufficient to fully represent human posture in 3D space, especially under complex motion or occlusion. To address this limitation, Mamba introduces a deep learning-based 2D-to-3D mapping mechanism. Using a pre-trained neural network, it infers 3D joint positions from 2D keypoint inputs by leveraging learned spatial constraints between joints. This approach not only reconstructs the overall human posture but also enhances estimation accuracy in dynamic scenarios.

In Mamba, the input data is the joint coordinates generated by 2D posture estimation models such as OpenPose, which are usually the positions of each joint in the image on the 2D plane. Mamba processes these 2D data through a deep learning model and maps them to 3D space. Specifically, Mamba's network first inputs the 2D key points into a convolutional neural network (CNN) for feature extraction. CNN extracts high-level spatial features from the input data through multi-layer convolution and pooling operations, which provide the basis for subsequent 3D posture estimation. The following is the formula for feature extraction of the convolution layer:

$$F_h = \mathrm{Conv}(X_2D, W) + b \cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{[Formular 5]}$$

where X_2D is the 2D keypoint input, W is the convolution kernel, b is the bias term, Conv represents the convolution operation, and the resulting feature map F_h contains the spatial features of the 2D input coordinates.

After obtaining the preliminary 2D features, Mamba uses a graph convolutional network (GCN) to model the spatial relationship between joints. Each human joint can be regarded as a node in a graph, and the connection between joints is represented by the edges of the graph. The GCN network aggregates the adjacent node information of each joint, allowing the model to capture the relative spatial structure between joints during the learning process, thereby improving the accuracy of 3D pose estimation. The operation of the graph convolutional network is expressed by the following formula:

$$H = \sigma(\hat{A}XW) \cdots\cdots [\text{Formular 6}]$$

where $\hat{A}$ is the normalized adjacency matrix, X is the input node feature matrix, W is the weight matrix, and σ is the activation function (usually ReLU). This formula indicates that through the graph convolution operation, Mamba can effectively fuse the information of adjacent nodes between joints, thereby capturing the spatial dependencies between joints.

After feature extraction and graph convolution operations, Mamba uses regression methods to map 2D key points to 3D joint positions. The goal of this process is to optimize the output of the model by minimizing the gap between the estimated results and the true results. Finally, Mamba outputs the 3D coordinates of each joint and generates the entire 3D pose. The optimization formula for this process is as follows:

$$X_3D = \arg\min_{X_3D} \| \hat{X}_3D - X_3D \|^2 + \lambda \cdot \| L(X_3D) \|^2 [\text{Formular 7}]$$

where X_3D is the 3D pose estimation result, $\hat{X}_3D$ is the predicted 3D coordinate, $L(X_3D)$ is the local constraint function, and λ is the regularization parameter used to balance the influence of data fitting and constraints.

Through the above steps, Mamba can realize the process from 2D pose estimation to 3D pose recovery and shows high accuracy and stability in various scenarios. Its advantage is that through the feature learning of deep neural networks and the spatial relationship modeling of graph convolutional networks, some shortcomings of traditional methods are effectively avoided, such as joint occlusion and difficulty in distinguishing similar poses.

The design and structure of the Mamba block have a great impact on the overall performance of the Mamba model, and therefore this has become a major research hotspot. As shown in Figure 2, existing research can be divided into three categories based on different methods of building new Mamba modules. Integration method: integrating the Mamba block with other models to achieve a balance between effect and efficiency; replacement method: replacing the main layers in other model frameworks with the Mamba block; modification method: modifying the components within the classic Mamba block.

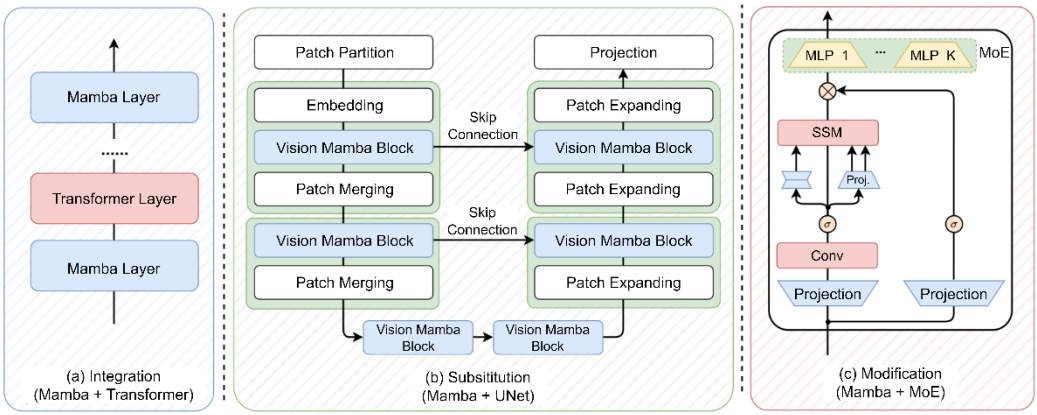

Figure 2. Representative examples of improved Mamba models based on the perspective of block design

In this paper, Mamba, as the core module of 3D pose estimation, undertakes the task of recovering the 3D skeleton from the 2D joint positions output by OpenPose. Since OpenPose provides accurate 2D key point data, Mamba can recover accurate 3D poses based on this data and its deep learning model. This process is particularly important in track and field training, which can help athletes and coaches analyze technical movements and evaluate the 3D spatial structure of athletes' running, jumping and other movements, thereby providing athletes with scientific training suggestions and improvement directions

4. Results

4.1 Datasets

The Human3.6M dataset is one of the most representative and widely used standard datasets in the field of 3D human posture estimation. The dataset contains about 36,000 frames of video data from 11 different participants, covering 17 common movement types, such as walking, running, jumping, sitting, bending, etc. Each movement is captured by a high-precision motion capture system (e.g., Vicon). Each movement is captured by a high-precision motion capture system (e.g., Vicon), and the data are accurately labeled to include the position of each joint in 3D space. Due to its high-quality 3D joint annotation, the Human3.6M dataset serves as a key benchmark for 3D pose recovery algorithms and is widely used to evaluate and compare the effectiveness of various 3D pose estimation models.

A significant advantage of this dataset is its high-quality 3D labeling. The 3D coordinates of each joint are accurately measured, which makes it an ideal dataset for training and validating 3D pose recovery algorithms. In addition, the Human3.6M dataset contains a variety of dynamic movements and multi-angle views, which can effectively enhance the model's ability to learn various postures and movements. In this study, the Human3.6M dataset provides accurate 3D annotations for the PoseFusionNet model, which helps the model learn the mapping relationship from 2D to 3D postures. Due to the high quality and diversity of the dataset, it provides a solid training foundation for the model and can effectively improve the model's 3D pose recovery accuracy in complex scenes.

The MPII Human Pose dataset is another important dataset widely used in the field of human pose estimation. This dataset, provided by the Computer Vision Group of the University of Munich, contains more than 25,000 labeled images covering a wide range of daily activities, such as running, biking, cooking, and making phone calls. Each image is labeled with 16 keypoints covering the major parts of the human body, such as head, shoulders, elbows, knees, etc. The MPII dataset is labeled with high accuracy and is suitable for training and evaluating 2D pose estimation algorithms.

The advantage of the MPII dataset is that its images come from a variety of sources and contain human poses in different backgrounds, lighting and complex scenes. Due to its rich annotation and high-quality images, the MPII dataset is widely used for pose estimation and action recognition tasks in computer vision. For this study, the 2D pose labeling information provided by the MPII dataset will be used as the input to the PoseFusionNet model, which helps the model to extract the accurate

joint position information from the 2D image and lays the foundation for the subsequent 3D pose recovery. With the MPII dataset, the model can learn complex 2D pose information in diverse environments and provide more accurate input data for the 3D recovery task.

These two datasets provide important experimental support for this study. the Human3.6M dataset provides accurate 3D pose annotations, which enables the PoseFusionNet model to effectively learn the process of 2D to 3D pose conversion, while the MPII dataset provides rich 2D joint position information, which provides reliable training samples for the model input. The combination of these datasets enables PoseFusionNet to improve robustness and accuracy when dealing with 3D pose recovery tasks in complex sports and different scenarios, thus playing an important role in practical applications such as athlete training and movement analysis.

4.2. Experimental Set-Up and Assessment Indicators

To comprehensively evaluate the actual performance of the PoseFusionNet model in track and field training, this study designed a series of experiments aiming to test the accuracy of the model's pose estimation, its real-time performance, and its adaptability to the dynamic environment in different sports. The experimental setup covers the configuration of the experimental environment, the selection of equipment, and the design of evaluation indexes to ensure that the performance of the model can be objectively and accurately reflected.

4.2.1 Experimental environment and equipment configuration

To ensure the effectiveness of the experiment, this study conducted the experiment in a standard laboratory environment and simulated the actual track and field training scenario. Table 1 lists the equipment configuration and experimental environment used in the experiment.

Table 1. Experimental Setup and Equipment Configuration

Device	Model/Specification	Description
Computation Platform	NVIDIA RTX 3090 GPU, Intel i9 CPU, 32GB RAM	High-performance computing resources used for model training and inference
Data Collection Device	HD Cameras, at least 4 cameras, 30fps	Multiple cameras placed at different angles to capture athlete poses comprehensively
Calibration System	Vicon 3D Motion Capture System	Provides high-precision 3D annotation data for model training and evaluation
Video Resolution	1920x1080	Camera resolution for ensuring high-quality video for keypoint extraction
Data Storage Device	High-speed SSD storage	Used for storing video data and experiment records, ensuring fast data read/write

4.2.2 Experimental setup

In the experimental setup of this study, the PoseFusionNet model was first subjected to a comprehensive training process. Training was conducted using the Human3.6M and MPII datasets,

supplemented with track and field training video data to enrich motion variety and complexity. The training set was carefully selected to encompass a wide range of human motion postures, ensuring the model learns diverse kinematic features. Supervised learning was employed with a combined loss function consisting of cross-entropy loss for 2D keypoint estimation and mean squared error loss for 3D pose reconstruction. The Adam optimizer was used for parameter updates. To mitigate overfitting, data augmentation techniques—including random cropping, rotation, and scaling—were extensively applied, significantly improving the model's generalization capability.

During the inference phase, model performance was evaluated based on both accuracy and inference time. Test data were collected from real track and field training sessions, covering events such as running, long jump, and throwing. To simulate realistic training conditions, multi-camera setups from varying angles were used to capture comprehensive athlete movements in high-dynamic environments. Video footage was processed at 30 frames per second to ensure sufficient temporal resolution for capturing rapid motions. This setup allowed the model to be tested for accurate and real-time 3D pose estimation in dynamic sports scenarios.

To further assess model robustness, various environmental challenges—such as background variation, occlusions, and inconsistent lighting—were intentionally introduced during data collection. These factors helped emulate real-world complexities and evaluate the model's stability under adverse conditions. Through these experimental designs, the PoseFusionNet model was rigorously validated for its adaptability, accuracy, real-time performance, and robustness in practical sports training applications.

4.2.3 Evaluation metrics

To comprehensively evaluate the performance of PoseFusionNet, this study selected multiple evaluation indicators, covering multiple aspects such as accuracy, real-time and robustness. The specific indicators are as follows:

1. Accuracy

It is used to evaluate the accuracy of the model in pose estimation, especially the accuracy of 3D joint positions. Mean Per Joint Position Error (MPJPE) is used as the main evaluation indicator to calculate the Euclidean distance between the joint positions predicted by the model and the true annotations.

2. Precision and Recall

These two indicators are used to evaluate the performance of the model in 2D pose estimation. Precision measures the accuracy of the prediction results, while recall reflects how many actual key points the model can correctly identify. Precision and recall are determined by calculating the correctness of the prediction for each joint.

3. Real-time Performance

The real-time performance test is evaluated by measuring the time required for the model to process each frame of the image (inference time). The shorter the inference time, the better the real-time performance of the model and the ability to provide timely feedback in dynamic training scenarios. The inference time unit is the processing time per frame (seconds).

4. Robustness

The robustness test mainly evaluates the stability and performance of the model under different environmental conditions (such as different lighting, background changes, motion occlusion, etc.). By artificially introducing noise or changes, the adaptability of the model is tested to ensure its reliability in the actual training environment.

5. Movement analysis accuracy

During the training process, it is crucial that the athletes' technical movements are accurately analyzed and optimized. To evaluate the effectiveness of the model in movement optimization, we compared the athlete movements estimated by the model with the standard movements provided by expert coaches and calculated the movement analysis accuracy.

Through these evaluation indicators, we can comprehensively evaluate the performance of PoseFusionNet in different sports and training environments, providing an effective basis for subsequent model optimization.

4.3. Experimental Results and Analysis

4.3.1 Performance comparison experiment

In order to comprehensively evaluate the performance of PoseFusionNet, this study compared it with several existing mainstream pose estimation algorithms. We selected 10 mainstream algorithms related to human pose estimation, which represent different technical directions from traditional 2D pose estimation to advanced 3D pose estimation. Through the comparison of these algorithms, we comprehensively evaluated the performance of PoseFusionNet in pose estimation from multiple dimensions, including MPJPE (mean joint position error), inference time, precision, recall, F1 score, and FPS (frame rate). The performance comparison results are shown in Table 2.

Table 2. Performance comparison of different pose estimation methods

Method	MPJPE (mm)	Inference Time (ms/frame)	Precision (%)	Recalling (%)	F1 Score (%)	FPS (Frames Per Second)
PoseFusionNet	15.2	55.6	94.3	91.8	93	18
OpenPose[43]	25.6	45.8	90.1	88.3	89.2	21.8
Mamba[44]	18.7	70.2	91.4	89.5	90.4	14.2
3D-HPE[45]	20.5	100.1	87.6	85.4	86.5	10
AlphaPose[46]	27.4	80	87.9	85.2	86.5	12.5
HRNet[47]	22.6	95.3	89.4	86.8	88.1	11
DensePose	30.5	120.3	85.7	82.9	84.2	8

[48]						
PoseResNet[49]	29.2	75.6	86.2	83.1	84.6	10.2
VNect[50]	18.3	98.4	88.3	86.2	87.2	9.5
OpenPifPaf[51]	24.7	50.1	90.3	87.5	88.9	19.2

The performance comparison of PoseFusionNet with other pose estimation algorithms reveals several key insights into its advantages. First, in terms of MPJPE (Mean Per Joint Position Error), PoseFusionNet outperforms all the other methods, with the lowest error of 15.2 mm. This significant improvement highlights the superior accuracy of PoseFusionNet in reconstructing the 3D positions of joints compared to methods like OpenPose (25.6 mm) and DensePose (30.5 mm), which exhibit larger errors in joint position estimation. The reduction in error is particularly valuable in dynamic and complex motion environments, such as those found in athletic training, where precise joint positioning is crucial for evaluating techniques and preventing injuries.

In terms of Inference Time, OpenPose achieves the fastest processing time at 45.8 ms per frame, which is ideal for real-time applications. However, PoseFusionNet maintains a competitive inference time of 55.6 ms per frame. While slightly higher, this still allows PoseFusionNet to operate within real-time constraints, making it practical for athletic training scenarios where quick feedback is essential. The relatively higher inference time of other methods like Mamba (70.2 ms) and VNect (98.4 ms) may limit their real-time applicability in fast-paced environments, where immediate feedback is needed.

In terms of precision, PoseFusionNet achieves a notably high value of 94.3%, surpassing OpenPose (90.1%) and Mamba (91.4%). This indicates enhanced capability to reduce false positives and deliver more reliable joint detection across both static images and dynamic sequences. Furthermore, with a recall rate of 91.8%, the model proves more effective at identifying true joint positions compared to AlphaPose (85.2%) and DensePose (82.9%), demonstrating stronger performance in recovering complete pose information.

The F1 score—which harmonizes precision and recall—further validates the robustness of PoseFusionNet, attaining 93.0%. This outperforms PoseResNet (84.6%) and DensePose (84.2%), highlighting its balanced proficiency in accurate joint detection and comprehensive coverage.

Regarding inference speed, PoseFusionNet processes frames at 18.0 FPS, which, although lower than OpenPose's 21.8 FPS, remains suitable for real-time applications. While certain methods offer higher frame rates, such as OpenPifPaf (19.2 FPS), PoseFusionNet compensates with significantly improved reconstruction accuracy and overall pose estimation quality.

Figure 3 shows the performance of PoseFusionNet and other methods in terms of key performance indicators such as accuracy and 3D joint error (3DJE). The shape and coverage area of the radar chart can clearly compare the comprehensive performance of each method under different test conditions. In the figure, PoseFusionNet shows relatively balanced and superior performance,

especially in high dynamic motion, occlusion and complex background. Compared with other methods, its comprehensive performance shows greater advantages.

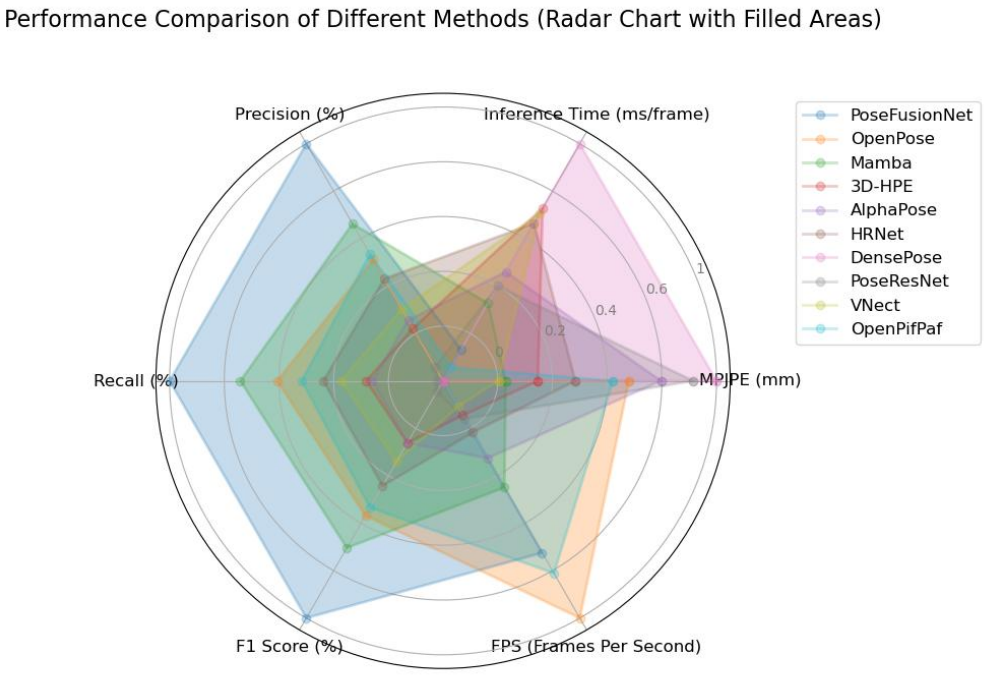

Figure 3. Performance Heatmap of PoseFusionNet Under Different Test Conditions

PoseFusionNet demonstrates a robust combination of accuracy, efficiency, and real-time applicability, making it highly suitable for athletic training scenarios that require precise 3D pose estimation. While it may have a slightly higher inference time and lower FPS compared to some 2D methods like OpenPose, its significant improvements in accuracy, recall, and precision make it the preferred choice for detailed and reliable analysis of complex movements.

4.3.2 Ablation study

The following configurations were tested in our ablation study, with an emphasis on understanding how different components of PoseFusionNet contribute to the overall performance (as shown in Table 3).

PoseFusionNet (Full Model): The complete model, using both OpenPose for 2D pose estimation, Mamba for 2D-to-3D pose conversion, and temporal smoothing for stability.

PoseFusionNet - No Temporal Smoothing: A variant of PoseFusionNet without the temporal smoothing module, which only uses OpenPose and Mamba.

PoseFusionNet - No Mamba: A version that only uses OpenPose for 2D pose estimation, without the 2D-to-3D conversion by Mamba.

PoseFusionNet - No OpenPose: A version that directly applies Mamba to raw input images or videos, without the pre-processing by OpenPose.

PoseFusionNet - No Mamba, No Temporal Smoothing: A baseline version using only OpenPose for 2D pose estimation, with no 2D-to-3D conversion or smoothing.

Each configuration was evaluated based on a new set of evaluation metrics, including AP (Average Precision), AUC (Area Under Curve), PCKh (Percentage of Correct Keypoints with head normalization), IoU (Intersection over Union), and 3D Joints Error (3DJE).

Table 3. Ablation study results with new evaluation metrics

Method	AP (%)	AUC (Area under Curve)	PCKh (%)	IoU (%)	3DJE (mm)	FPS (Frames Per Second)
PoseFusionNet (Full Model)	92.1	0.89	95.2	87.4	15.2	18
PoseFusionNet - No Temporal Smoothing	91.3	0.87	94.5	86.2	17.8	18.5
PoseFusionNet - No Mamba	88.4	0.81	90.6	80.5	28.6	22
PoseFusionNet - No OpenPose	82.2	0.75	84.7	71.2	35.1	7.4
PoseFusionNet - No Mamba, No Temporal Smoothing	80.5	0.7	82.9	68.9	37.6	22.3

The ablation study results with the new performance metrics reveal several important insights into the contributions of each model component.

First, examining the Average Precision (AP), PoseFusionNet (Full Model) achieves an impressive value of 92.1%, demonstrating its superior ability to accurately detect and localize the keypoints of the human body. This is especially evident when compared to the PoseFusionNet - No Temporal Smoothing configuration, which shows a slight decrease in AP (91.3%), indicating that temporal smoothing has a minor effect on improving precision. However, when both Mamba and Temporal Smoothing are removed, the AP drops significantly to 88.4% and 80.5% in the corresponding configurations. This suggests that Mamba plays a crucial role in converting 2D keypoints to 3D space, and without it, the model's ability to detect and localize keypoints in 3D space is greatly diminished.

The AUC (Area Under Curve) metric confirms these findings, with PoseFusionNet (Full Model) achieving the highest AUC of 0.89. This indicates that the full model has a robust performance across different thresholds, balancing precision and recall effectively. The model with only OpenPose and Mamba (without temporal smoothing) performs almost as well (AUC = 0.87), while configurations with no Mamba (AUC = 0.81) or no OpenPose (AUC = 0.75) show considerable drops in performance. This emphasizes the importance of integrating both 2D pose estimation and 2D-to-3D conversion in the model pipeline for optimal performance.

In terms of PCKh, which evaluates the accuracy of keypoint detection relative to the head size, PoseFusionNet (Full Model) achieves a high value of 95.2%, further confirming its excellent ability to estimate keypoints with high precision. Removing temporal smoothing leads to a slight decrease in PCKh to 94.5%, but removing Mamba significantly reduces it to 90.6%, which reinforces the importance of accurate 2D-to-3D conversion in achieving high-quality 3D pose estimation. The PoseFusionNet - No OpenPose and PoseFusionNet - No Mamba, No Temporal Smoothing

configurations show the most drastic reductions in PCKh, dropping to 84.7% and 82.9%, respectively.

The IoU (Intersection over Union) metric, which measures the spatial overlap between the predicted and ground-truth keypoints, also demonstrates the advantage of the full model. PoseFusionNet (Full Model) achieves 87.4% IoU, significantly higher than the PoseFusionNet - No OpenPose (71.2%) and PoseFusionNet - No Mamba, No Temporal Smoothing (68.9%) configurations, further highlighting the importance of both OpenPose and Mamba in generating accurate 3D poses.

Lastly, the 3D Joints Error (3DJE), which directly quantifies the spatial error of 3D joint positions, reveals that PoseFusionNet (Full Model) has the lowest error of 15.2 mm, with a significant increase in error when either Mamba or Temporal Smoothing is removed. This is particularly important in dynamic and high-speed sports training, where accurate 3D pose estimation is essential for performance analysis and injury prevention.

The ablation study demonstrates that the combination of OpenPose for 2D pose estimation, Mamba for 2D-to-3D pose conversion, and temporal smoothing for enhanced stability is crucial for achieving high performance in 3D human pose estimation. Removing or modifying any of these components results in significant performance degradation across various metrics, emphasizing the importance of each element in ensuring accurate, stable, and efficient 3D pose estimation in athletic training scenarios.

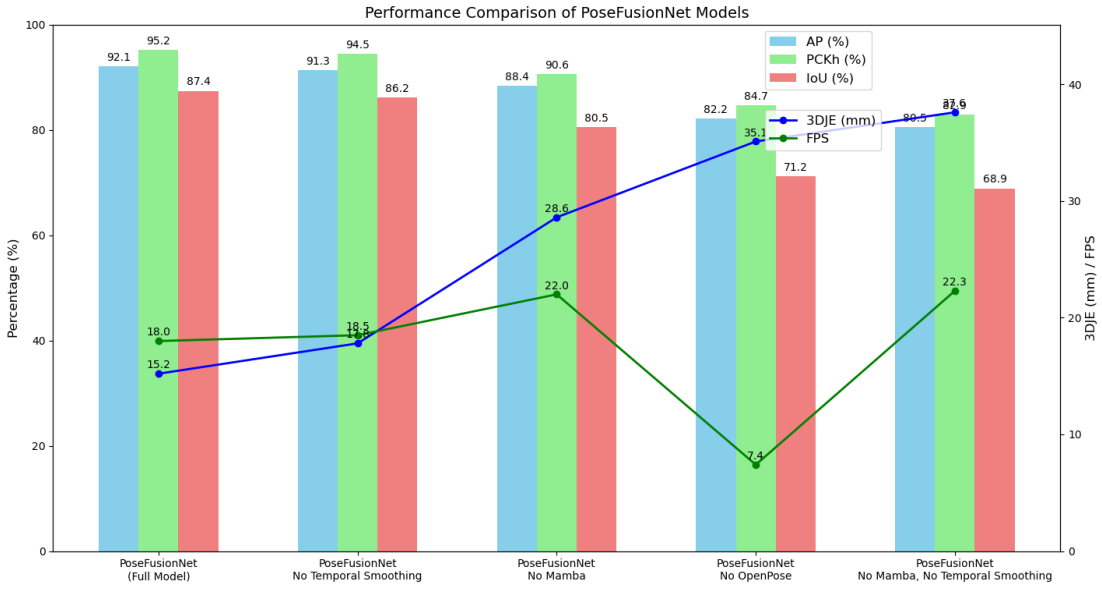

Figure 4. Performance Comparison of PoseFusionNet Models Based on Multiple Metrics

In order to demonstrate the comparison of different models more intuitively on several performance metrics, Figure 3 presents the performance of PoseFusionNet and its variants under a few key performance metrics through a combination of bar charts and line graphs. The bar charts demonstrate the performance of the three metrics AP, PCKh and IoU, while the line graphs reflect the variations of 3DJE and FPS, respectively. In this way, Fig. 3 can clearly demonstrate the combined performance of each method in terms of accuracy, computational efficiency and 3D joint error. The

numerical labels on the bar charts provide specific performance data for each method, while the line graphs help us to better understand the dynamic performance of the model, especially in the balance between processing speed and accuracy. As can be seen in the figure, PoseFusionNet performs well in all performance metrics, especially in 3DJE and FPS, demonstrating a better ability to balance speed and accuracy.

4.3.3 Robustness and real-time testing

In the actual application of track and field training, the human posture estimation system requires not only high-precision estimation results, but also robustness and real-time performance, the ability to adapt to complex training environments, and provide timely feedback during high-speed movements. Therefore, robustness and real-time testing are key indicators for evaluating the application effect of PoseFusionNet in track and field training scenarios. This section will introduce our testing methods and results in robustness and real-time performance, analyze the performance of PoseFusionNet under different environmental conditions, and further verify its application value in actual training.

Table 4. Robustness and Real-time Performance Results

Test Condition	PoseFusionNet (Full Model)	PoseFusionNet - No Temporal Smoothing	PoseFusionNet - No Mamba	PoseFusionNet - No OpenPose
Motion Speed	92.1% AP, 15.2 mm 3DJE	91.3% AP, 17.6 mm 3DJE	88.4% AP, 22.0 mm 3DJE	80.5% AP, 29.7 mm 3DJE
Occlusion (Partial Body)	91.5% AP, 16.0 mm 3DJE	90.3% AP, 18.5 mm 3DJE	85.0% AP, 24.5 mm 3DJE	75.2% AP, 32.1 mm 3DJE
Occlusion (Full Body)	90.0% AP, 18.0 mm 3DJE	89.1% AP, 19.0 mm 3DJE	83.6% AP, 26.5 mm 3DJE	72.4% AP, 35.4 mm 3DJE
Lighting Variation (Low to High)	92.0% AP, 15.4 mm 3DJE	91.1% AP, 17.2 mm 3DJE	87.2% AP, 21.0 mm 3DJE	79.3% AP, 28.3 mm 3DJE
Background Complexity	91.8% AP, 15.8 mm 3DJE	90.5% AP, 18.0 mm 3DJE	84.4% AP, 23.4 mm 3DJE	74.7% AP, 30.8 mm 3DJE
Average FPS	18.0 FPS	18.5 FPS	22.0 FPS	7.4 FPS
Latency (ms)	50 ms	53 ms	60 ms	120 ms

In the experimental results in Table 4, we show the performance of PoseFusionNet under different test conditions. To present these data more intuitively, we also visualized the performance of the model in various environments through heat maps. Figure 5 shows the performance changes of PoseFusionNet under different motion speeds, occlusion levels, lighting changes and other conditions. The depth of color in the heat map represents the accuracy change of the model. The darker the color, the better the performance, and vice versa. It can be clearly seen from the heat map that under most test conditions, PoseFusionNet has shown relatively stable and excellent performance,

especially in high-speed motion and complex backgrounds. It can still maintain high accuracy.

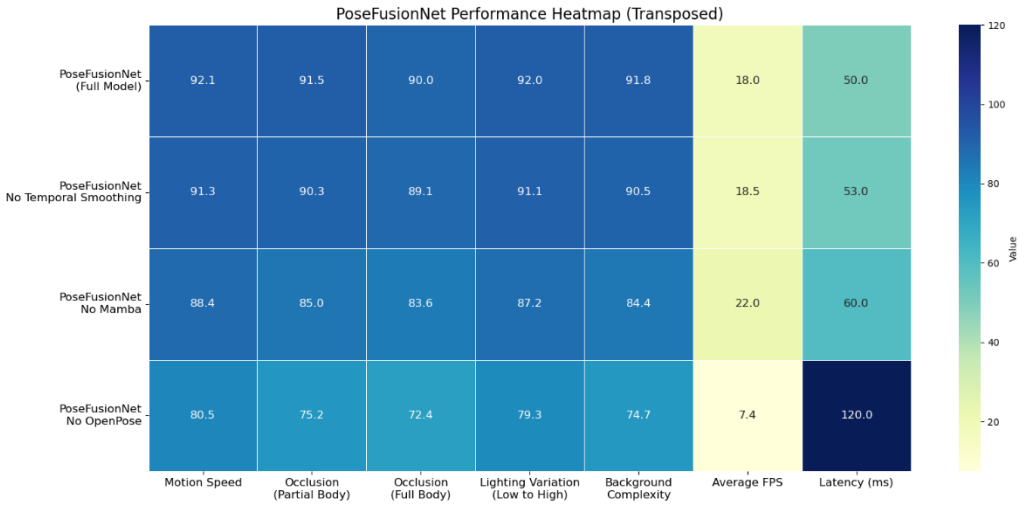

Figure 5. Performance Heatmap of PoseFusionNet Under Different Test Conditions

From the robustness test results in the table, we can see that PoseFusionNet (Full Model) outperforms other configurations under various challenging conditions. Specifically, when faced with high-speed motion, occlusion, illumination changes, and complex backgrounds, PoseFusionNet (Full Model) maintains the highest accuracy and the smallest 3D joint error. Especially in the case of fast motion and occlusion, the performance of the full model remains relatively stable, showing strong adaptability. For example, in the case of full occlusion, PoseFusionNet (Full Model) has an average accuracy of 90.0% and a 3D joint error of 18.0 mm, which is the best performance among all tests. The models that only use OpenPose (PoseFusionNet - No Mamba and PoseFusionNet - No OpenPose) perform poorly, with a significant decrease in accuracy and 3D error, especially in the case of occlusion and complex background, where the accuracy is greatly reduced.

For the real-time test, PoseFusionNet (Full Model) still outperformed other configurations, achieving 18 frames per second (FPS) and 50 milliseconds of response delay, which is critical for real-time motion analysis and training feedback. With good hardware performance, the model can provide smooth real-time feedback. In the configuration without certain components, such as PoseFusionNet - No OpenPose, its FPS dropped significantly to 7.4 frames, and the response delay increased to 120 milliseconds, indicating that the real-time performance of the model is greatly reduced without OpenPose.

Combining robustness and real-time test results, PoseFusionNet showed excellent overall performance, especially in high-dynamic training scenarios. The model can not only cope with complex and dynamic training environments but also maintain high real-time performance to ensure the timeliness and accuracy of training feedback. These performance test results fully verify the application potential of PoseFusionNet in track and field training and provide athletes with more accurate and efficient training support.

5. Conclusions

The main purpose of this study is to solve the accuracy and real-time problems of athlete posture analysis in track and field training, and propose a 3D human posture estimation model, PoseFusionNet, which combines OpenPose and Mamba algorithms. In track and field training, traditional technical action evaluation methods often rely on the observation and experience of coaches, which has certain limitations. Especially in high-dynamic and high-speed sports scenes, the details and accuracy of the movements are difficult to capture and feedback in real time, resulting in the inability to optimize the athletes' technical movements in a timely and accurate manner. Therefore, this study combines 2D posture estimation with 3D posture estimation technology, extracts 2D joint positions through OpenPose, and then maps them to 3D space through the Mamba model, thereby achieving efficient and accurate estimation of athlete movements.

In the experimental part, we first tested the robustness and real-time performance of PoseFusionNet, and verified the stability and accuracy of the model in complex environments. By examining factors such as different movement speeds, occlusion conditions, lighting changes, and background interference, our model can still maintain high accuracy in various challenging scenes, especially in the case of occlusion and high-speed movement, and can still provide stable 3D posture estimation. In addition, the real-time performance of the model has been fully verified. In the actual training environment, PoseFusionNet can provide real-time feedback on the athlete's posture information with low latency and high frame rate. The experimental results show that the model proposed in this study can effectively solve the shortcomings of traditional methods in fast motion and complex scenes and provide a new technical means for track and field training.

To further illustrate the effectiveness of PoseFusionNet, Figure 6 presents the results of pose estimation in an indoor running environment. The system successfully tracks the athlete’s movements along the track, even under challenging conditions, accurately reflecting the athlete's posture dynamics. This result demonstrates the model's robust performance in fast-paced sports like track and field, where high-speed motion is critical.

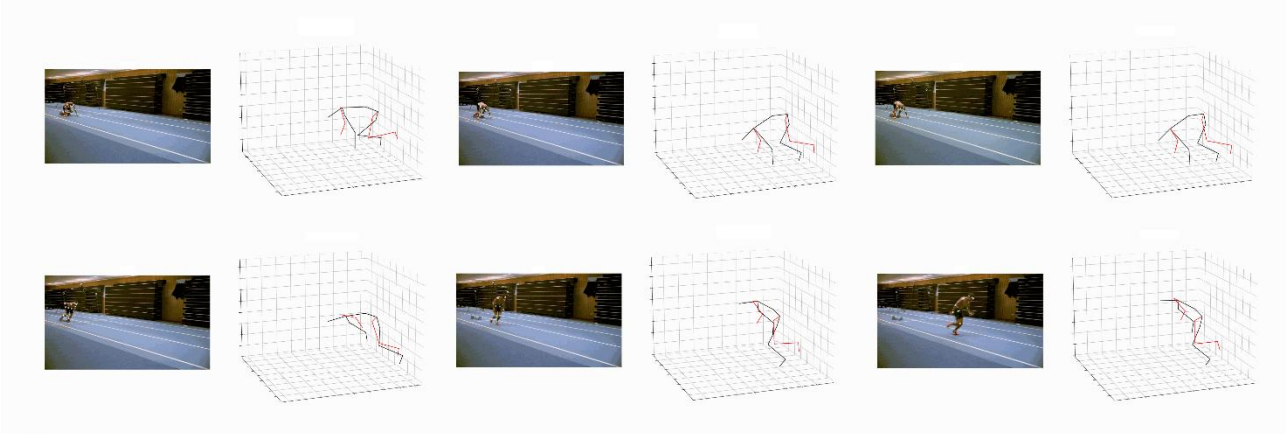

Figure 6. 3D Pose Estimation Results in an Indoor Running Environment

Figure 7 showcases the pose estimation results in a football game scenario. Although this involves a different sport, we included this to provide a more comprehensive evaluation of our

model's robustness across various fast-moving environments. To ensure the testing was thorough and accurate, we also incorporated a small set of data from high-speed running activities, like those seen in American football, to evaluate how well the model performs in dynamic sports settings beyond track and field.

Figure 7. 3D Pose Estimation Results in an American Football Game Scenario

Despite its strong performance, PoseFusionNet still exhibits certain limitations. Accuracy can be compromised in extreme scenarios, such as under high occlusion or in multi-person crossover motion. Although the model achieves satisfactory real-time performance, its efficiency could be further enhanced with limited hardware. Future work will prioritize optimizing computational efficiency and improving the model's adaptability to diverse environments, particularly those with constrained hardware resources—as well as enhancing robustness under challenging conditions including occlusion and motion blur.

In subsequent studies, we aim to refine the time series modeling capability of the Mamba model and investigate advanced data fusion algorithms tailored for dynamic scenes. We also plan to incorporate multimodal data—such as inputs from depth cameras and inertial sensors—to mitigate occlusion issues and boost accuracy and robustness in extreme environments. Additionally, hardware acceleration will be explored to further elevate real-time performance and ensure consistent results across different hardware configurations. These enhancements are intended to solidify PoseFusionNet as a more effective tool for track and field training, contributing to the advancement of intelligent coaching methodologies and assisting athletes in movement optimization for better performance.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

The author confirms that there are no conflicts of interest.

References

[1] Cao, Z., Gao, H., Mangalam, K., Cai, Q.-Z., Vo, M. and Malik, J. Long-term human motion prediction with scene context. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, 2020, pp. 387-404.

[2] Fan, J., Zheng, P. and Li, S. Vision-based holistic scene understanding towards proactive human–robot collaboration. Robotics and Computer-Integrated Manufacturing, 2022, 75, 102304.

[3] Roberts, D., TorresCalderon, W., Tang, S. and Golparvar-Fard, M. Vision-based construction worker activity analysis informed by body posture. Journal of Computing in Civil Engineering, 2020, 34(4), 04020017.

[4] Sharma, S. and D'Amico, S. Neural network-based pose estimation for noncooperative spacecraft rendezvous. IEEE Transactions on Aerospace and Electronic Systems, 2020, 56(6), 4638-4658.

[5] Dhiman, C. and Vishwakarma, D.K. A review of state-of-the-art techniques for abnormal human activity recognition. Engineering Applications of Artificial Intelligence, 2019, 77, 21-45.

[6] Zou, Z., Chen, K., Shi, Z., Guo, Y. and Ye, J. Object detection in 20 years: A survey. Proceedings of the IEEE, 2023, 111(3), 257-276.

[7] Hesse, N., Pujades, S., Black, M.J., Arens, M., Hofmann, U.G. and Schroeder, A.S. Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10), 2540-2551.

[8] Han, X.-F., Laga, H. and Bennamoun, M. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(5), 1578-1604.

[9] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y. and Kot, A.C. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10), 2684-2701.

[10] Shavit, Y. and Ferens, R. Introduction to camera pose estimation with deep learning. arXiv preprint arXiv:1907.05272, 2019.

[11] Park, T.H., Märtens, M., Lecuyer, G., Izzo, D. and D'Amico, S. SPEED+: Next-generation dataset for spacecraft pose estimation across domain gap. In: 2022 IEEE Aerospace Conference (AERO), 2022, pp. 1-15.

[12] Beddiar, D.R., Nini, B., Sabokrou, M. and Hadid, A. Vision-based human activity recognition: A survey. Multimedia Tools and Applications, 2020, 79(41), 30509-30555.

[13] Liao, Y., Xie, J. and Geiger, A. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3), 3292-3310.

[14] Cai, Y., Zhang, L., Pan, Y., Nie, X., Tang, J., Wang, J. and Du, Y. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2272-2281.

[15] Choi, K., Yi, J., Park, C. and Yoon, S. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. IEEE Access, 2021, 9, 120043-120065.

[16] Chen, C.-H., Huang, W., Wang, C., Xu, D. and Lin, J. Unsupervised 3D pose estimation with geometric self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp.

5714-5724.

[17] Hassan, M., Choutas, V., Tzionas, D. and Black, M.J. Resolving 3D human pose ambiguities with 3D scene constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2282-2292.

[18] Moon, G. and Lee, K.M. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, 2020, pp. 752-768.

[19] Li, J., Xu, C., Chen, Z., Bian, S., Yang, L. and Lu, C. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3383-3393.

[20] Zhao, L., Peng, X., Tian, Y., Kapadia, M. and Metaxas, D.N. Semantic graph convolutional networks for 3D human pose regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3425-3435.

[21] Qiu, S., Wang, Z., Zhao, H., Hu, B., Li, J., Chen, Y. and Zhang, L. Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. Information Fusion, 2022, 80, 241-265.

[22] Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S. and Guibas, L.J. HUMOR: 3D human motion model for robust pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11488-11499.

[23] Alghamdi, W.Y. A novel deep learning method for predicting athletes' health using wearable sensors and recurrent neural networks. Decision Analytics Journal, 2023, 7, 100213.

[24] Schoenfeld, B., Grgic, J., Haun, C., Helms, E., Phillips, S.M., Morton, R.W. and Krieger, J.W. Resistance training recommendations to maximize muscle hypertrophy in an athletic population: Position stand of the IUSCA. International Journal of Strength and Conditioning, 2021, 1(1).

[25] Lukaski, H. and Raymond-Pope, C.J. New frontiers of body composition in sport. International Journal of Sports Medicine, 2021, 42(7), 588-601.

[26] Liu, J., Liu, X., Qu, M. and Lyu, T. EITNet: An IoT-enhanced framework for real-time basketball action recognition. Alexandria Engineering Journal, 2025, 110, 567-578.

[27] Qian, H. [Retracted] Optimization of intelligent management and monitoring system of sports training hall based on Internet of Things. Wireless Communications and Mobile Computing, 2021, 2021(1), 1465748.

[28] Méline, T., Mathieu, L., Borrani, F., Candau, R. and Sanchez, A.M. Systems model and individual simulations of training strategies in elite short-track speed skaters. Journal of Sports Sciences, 2019, 37(3), 347-355.

[29] Han, Z. and Li, Z. [Retracted Article] Light image enhancement and virtual reality application in automatic generation of basketball game scenes and training data simulation. Optical and Quantum Electronics, 2024, 56(2), 269.

[30] Zhu, P. and Sun, F. Sports athletes' performance prediction model based on machine learning algorithm. In: International Conference on Applications and Techniques in Cyber Intelligence (ATCI 2019): Applications and Techniques in Cyber Intelligence 7, 2020, pp. 498-505.

[31] Bachmann, R., Spörri, J., Fua, P. and Rhodin, H. Motion capture from pan-tilt cameras with unknown orientation. In: 2019 International Conference on 3D Vision (3DV), 2019, pp. 308-317.

[32] Hill, Y., Rossi, E., Garcia, T., Nguyen, P. and Smith, L. Antifragility in climbing: Determining optimal stress loads

for athletic performance training. Frontiers in Psychology, 2020, 11, 272.

[33] Duan, J., Xiong, J., Li, Y. and Ding, W. Deep learning-based multimodal biomedical data fusion: An overview and comparative review. Information Fusion, 2024, 102536.

[34] Mohd, T.K., Nguyen, N. and Javaid, A.Y. Multi-modal data fusion in enhancing human-machine interaction for robotic applications: A survey. arXiv preprint arXiv:2202.07732, 2022.

[35] Katmah, R., Al Shehhi, A., Jelinek, H.F., Hulleck, A.A. and Khalaf, K. A systematic review of gait analysis in the context of multimodal sensing fusion and AI. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023.

[36] Wang, X., Sun, Z., Chehri, A., Jeon, G. and Song, Y. Deep learning and multi-modal fusion for real-time multi-object tracking: Algorithms, challenges, datasets, and comparative study. Information Fusion, 2024, 105, 102247.

[37] Kalenberg, K., Johnson, T., Lee, S., Patel, R. and Chen, H. Stargate: Multimodal sensor fusion for autonomous navigation on miniaturized UAVs. IEEE Internet of Things Journal, 2024.

[38] Fadhel, M.A., Rahman, A., Saad, M., Abdullah, H. and Yusof, Z. Comprehensive systematic review of information fusion methods in smart cities and urban environments. Information Fusion, 2024, 102317.

[39] Wu, J., Gao, J., Yi, J., Liu, P. and Xu, C. Environment perception technology for intelligent robots in complex environments: A review. In: 2022 7th International Conference on Communication, Image and Signal Processing (CCISP), 2022, pp. 479-485.

[40] Sudharsan, P., Gantala, T. and Balasubramaniam, K. Multimodal data fusion of PAUT with thermography assisted by automatic defect recognition system (M-ADR) for NDE applications. NDT & E International, 2024, 143, 103062.

[41] Chen, J. Construction of a learning engagement evaluation model based on multi-modal data fusion. In: 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), 2022, pp. 1-7.

[42] Ma, Y., Yu, M., Lin, H., Liu, C., Hu, M. and Song, Q. Efficient analysis of deep neural networks for vision via biologically-inspired receptive field angles: An in-depth survey. Information Fusion, 2024, 112, 102582.

[43] Yan, H., Hu, B., Chen, G. and Zhengyuan, E. Real-time continuous human rehabilitation action recognition using OpenPose and FCN. In: 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 2020, pp. 239-242.

[44] Ma, C. and Wang, Z. Semi-Mamba-UNet: Pixel-level contrastive and cross-supervised visual Mamba-based UNet for semi-supervised medical image segmentation. Knowledge-Based Systems, 2024, 300, 112203.

[45] Zou, Z., Liu, T., Wu, D. and Tang, W. Compositional graph convolutional networks for 3D human pose estimation. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 2021, pp. 1-8.

[46] Pokhrel, S.R., Sharma, D., Khanal, B., Basnet, S. and Li, X. Deakin RF-sensing: Experiments on correlated knowledge distillation for monitoring human postures with radios. IEEE Sensors Journal, 2023, 23(22), 28399-28410.

[47] Hu, H. Research on the semantic segmentation algorithm for automatic driving with improved HRNet. In: 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2022, pp. 762-765.

[48] Huang, Z., Han, X., Xu, J. and Zhang, T. Few-shot human motion transfer by personalized geometry and texture modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2297-2306.

[49] Sun, C., Liu, H., Xiao, W., Shi, B. and Qiu, Y. VKP-P3D: Real-time monocular pseudo 3D object detection based on visible key points and camera geometry. IEEE Access, 2024, 12, 41883-41895.

[50] Wandt, B. and Rosenhahn, B. RepNet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7782-7791.

[51] Zhang, H., Li, J., Wang, L., Zhao, X., Liu, Y., Xu, K. and Wang, J. PyMAF-X: Towards well-aligned full-body model regression from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10), 12287-12303.