PoseTrackNet: Robotic Vision-Driven Posture and Motion Analysis for Sports Training and Rehabilitation

*Dr. Rajesh Kumar KV

*Associate Professor, AI Research Centre, School of Business, Woxsen University, India; rajesh.kumar@woxsen.edu.in

*Corresponding Author: rajesh.kumar@woxsen.edu.in

DOI: https://doi.org/10.30211/JIC.202604.003

Submitted: Jan. 17, 2026 Accepted: Mar. 07, 2026

ABSTRACT

Robotic vision systems have become increasingly vital in sports training and rehabilitation, where precise posture and motion analysis is crucial for enhancing performance and reducing injury risks. However, current models often fall short in providing real-time corrective feedback, limiting their utility in dynamic and fast-paced environments. To address these challenges, we introduce PoseTrackNet, a novel integrated model that combines EfficientDet for object detection, OpenPose for pose estimation, and a Pose Correction Module designed to deliver immediate posture adjustments. This integration allows PoseTrackNet to offer superior accuracy, robust tracking, and low latency, making it ideal for real-world applications. In experiments conducted on the COCO and MPII Human Pose datasets, PoseTrackNet achieved a Mean Average Precision (mAP) of 43.7%, a Corrected Pose Percentage (CPP) of 87.3%, and an Error Reduction Rate (ERR) of 78.4%, significantly outperforming recent models. Additionally, the system maintains a Frame Processing Rate of 30 fps, ensuring it meets the demands of real-time applications. These results demonstrate PoseTrackNet's capability to provide precise, context-aware feedback, making it a powerful tool for both sports training and rehabilitation. This research not only addresses existing limitations in current systems but also sets a foundation for future developments in real-time robotic vision-driven posture and motion analysis.

Keywords: Real-time posture correction, Human pose estimation, Sports training analysis, Robotic vision, Rehabilitation technology

1. Introduction

Human pose estimation has become an essential technology across various domains, including sports training, physical rehabilitation, human-computer interaction, and augmented reality. The ability to accurately capture and analyze human movement in real-time offers considerable advantages, particularly in scenarios where correct posture and precise motion are critical for achieving optimal performance and minimizing the risk of injury. Over the past decade, advances in deep learning have significantly enhanced pose estimation capabilities. Notable models such as OpenPose have set new standards for detecting human keypoints with high precision, while EfficientDet has provided robust object detection across diverse contexts. However, despite these

advancements, a critical gap remains: while these systems excel at identifying and tracking poses, they do not inherently provide feedback or corrective guidance to address postural deviations as they occur. This limitation is particularly pronounced in dynamic and high-stakes environments like sports training, where real-time corrections are vital for both improving technique and preventing injuries.

Addressing this gap, this research introduces an integrated system that combines the pose estimation capabilities of OpenPose and the object detection strengths of EfficientDet with a novel Pose Correction Module. The primary objective of this study is to develop a system capable of not only detecting and tracking human poses but also actively evaluating these poses against predefined standards and providing real-time corrective feedback. This feedback loop enables the system to guide users in adjusting their movements during activities, ensuring that they maintain proper form throughout. This approach represents a significant shift from traditional pose estimation models, which are primarily passive and static in their analysis, offering no dynamic interaction or corrective measures for the user.

PoseTrackNet is technically novel in terms of notPoseTrackNet demonstrates technical novelty not only through the integration of EfficientDet and OpenPose, but also through the design of its Pose Correction Module and its tight coupling with these underlying models. Unlike prior systems that primarily treat pose estimation as a static analytical task, PoseTrackNet introduces three key advancements. First, it incorporates a biomechanics-informed library of ideal pose representations derived from expert-annotated templates for each target activity. Second, it employs a quantitative deviation assessment mechanism to systematically measure discrepancies between observed and ideal poses. Third, it features a multi-channel corrective feedback mechanism that translates these deviations into prioritized and actionable instructions across multiple feedback modalities.
Collectively, these components enable PoseTrackNet to function as an interactive, closed-loop posture coaching system, extending beyond conventional implementations that rely solely on pose estimation using EfficientDet and OpenPose.

The rationale behind integrating a Pose Correction Module lies in its potential to enhance the utility and impact of pose estimation systems across various applications. By leveraging advanced algorithms to detect and correct deviations from ideal postures, the module transforms a purely analytical tool into an interactive system capable of guiding users toward optimal performance in real-time. This research not only aims to bridge the current gap in pose estimation technology but also aspires to set a foundation for future developments in intelligent training systems. The integration of corrective feedback within pose estimation can extend beyond sports training, finding applications in rehabilitation, ergonomic assessments, and even daily fitness routines where maintaining correct posture is crucial.

In summary, the study seeks to develop a comprehensive system that unites the detection strengths of OpenPose and EfficientDet with the corrective capabilities of a newly developed Pose Correction Module. This system is designed to address the limitations of existing models by offering real-time, actionable feedback that enhances both the accuracy and effectiveness of human pose analysis. Through this research, we aim to contribute to the ongoing evolution of intelligent and adaptive systems, pushing the boundaries of what is possible in real-time posture monitoring and correction.

Our contributions are summarized as follows:

1. Integrated Pose Correction System: We developed a novel system that integrates EfficientDet and OpenPose with a Pose Correction Module, enabling real-time detection, analysis, and correction of human postures. This system provides dynamic feedback to users, improving posture accuracy during activities such as sports training and rehabilitation.
2. Enhanced Real-Time Feedback Mechanism: We introduced a real-time feedback loop that evaluates detected poses against predefined standards and issues corrective guidance instantaneously. This mechanism ensures that users maintain proper form throughout their movements, directly addressing the limitations of traditional pose estimation systems.
3. Comprehensive Evaluation and Improvement: We conducted extensive experiments to validate the effectiveness of the integrated system, demonstrating significant improvements in both posture accuracy and training outcomes. Our results highlight the system's potential to enhance user performance and reduce the risk of injury in various application scenarios.

In the following sections, we present a detailed exploration of our approach and findings. Section 2 reviews related work in the field of human pose estimation and pose correction, discussing recent advancements and identifying the key challenges that our research addresses. Section 3 outlines our methodology, describing the integration of EfficientDet, OpenPose, and the Pose Correction Module, and explaining how these components work together to provide real-time feedback and pose adjustment. Section 4 details the experimental setup, including the datasets used, evaluation metrics, and the environment in which the system was tested, followed by an analysis of the results and a comparison with existing methods. Finally, Section 5 concludes the paper by summarizing our contributions, discussing the broader implications of our findings, and suggesting directions for future research to further enhance real-time pose estimation and correction systems.

2. Literature Review

2.1 Current Status and Challenges of Computer Vision in Sports Training

The integration of computer vision technology into sports training has significantly transformed the way athletes' performance is monitored and evaluated [1]. By utilizing advanced image processing and pattern recognition algorithms, computer vision can automatically detect and track key points of the human body [2], facilitating detailed analysis of an athlete’s posture and movement patterns [3]. This technology has become a cornerstone for developing personalized training programs, enabling coaches and athletes to optimize performance with data-driven insights [4]. However, early computer vision methods, which relied heavily on manually engineered features and simplistic models [5], often struggled in real-world scenarios where the environment is dynamic and uncontrolled.

In recent years, the focus of research has shifted toward enhancing the robustness and real-time performance of pose estimation models [6]. Deep learning approaches, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have become prominent in extracting intricate spatiotemporal features from video data [7]. These models have proven effective in handling the complexities of human motion, such as occlusions [8], lighting variations, and background noise[9]. Additionally, innovative techniques like multi-task learning, adversarial learning, and cross-

modal data fusion are being actively explored to further improve the adaptability and precision of pose estimation models across diverse sports environments [10].

Despite these advancements, significant challenges remain. Current systems often falter in complex and fast-paced sports scenarios, where the accuracy of pose estimation can be compromised by factors such as rapid motion [11], background clutter, and variable lighting conditions. Moreover, existing methods frequently lack the capability to provide real-time feedback, which is crucial for interactive training applications [12]. This study addresses these limitations by introducing a novel approach that combines multiple models and algorithms, aiming to achieve higher efficiency and accuracy in posture analysis, even in challenging environments.

2.2 Advances in Deep Learning for Human Pose Estimation

The application of deep learning to human pose estimation has revolutionized the field, driving substantial improvements in accuracy and robustness [13]. Traditional methods, which relied on handcrafted features and heuristic algorithms [14], often fell short in generalizing to complex and diverse human poses [15]. With the advent of deep learning, models like OpenPose, HRNet, and more recently PoseFormer, have significantly enhanced the ability to detect and estimate human postures in both 2D and 3D spaces [16]. These models leverage the power of convolutional architectures to automatically learn and extract hierarchical features from large-scale datasets [17], enabling the recognition of fine-grained pose variations across different subjects and scenarios.

The current research landscape is increasingly focused on refining these models to address their inherent limitations. Key areas of interest include multi-scale feature extraction [18], where models like HRNet achieve superior accuracy by maintaining high-resolution representations throughout the network [19]. Another trend is the exploration of cross-domain transfer learning, which aims to adapt pose estimation models trained on one dataset to perform well on different datasets with minimal fine-tuning [20]. Additionally, self-supervised learning techniques are gaining traction as a means to reduce the dependency on large labeled datasets, which are often expensive and time-consuming to create [21]. Spatiotemporal modeling is also a major focus, where the integration of temporal dynamics with spatial features is critical for capturing the fluidity of human motion over time [6].

Nevertheless, the application of deep learning to pose estimation is not without its drawbacks. The high computational cost and the need for substantial processing power make it difficult to deploy these models in real-time scenarios [22], particularly in resource-constrained environments like mobile devices or embedded systems. Moreover, while these models excel in controlled settings, their performance often degrades in real-world situations involving complex backgrounds, occlusions, and diverse lighting conditions [23]. This study proposes a multi-model fusion approach that seeks to overcome these challenges by reducing computational complexity and enhancing the real-time capabilities of pose estimation systems, making them more suitable for practical sports training applications [24].

2.3 Exploration of Robotic Vision Technology in Sports Training

Robotic vision technology has increasingly been adopted in sports training, offering a sophisticated means of monitoring, and analyzing athletes' movements in real-time. By combining advanced imaging techniques with intelligent algorithms, robotic systems can provide instant

feedback on an athlete's performance [25], helping to correct posture and refine techniques. These systems are particularly valuable in high-intensity sports, where precise movement and form are critical to success [26]. The use of robotic vision allows for continuous monitoring, ensuring that athletes maintain proper form and avoid injury [27]. Applications range from posture correction in strength training to gait analysis in endurance sports, showcasing the versatility and effectiveness of these systems.

Recent advancements have focused on improving the accuracy, speed, and adaptability of robotic vision systems. Researchers are increasingly exploring the integration of multi-view and multimodal data, such as combining video feeds with data from wearable sensors, to enhance the depth and reliability of posture analysis. The development of adaptive algorithms that can adjust to different athletes and training conditions is also a key area of innovation. Additionally, the implementation of intelligent feedback mechanisms enables robots to provide personalized guidance in real-time, further enhancing the training experience and outcomes.

However, despite these advancements, there are still significant obstacles to overcome. One of the primary challenges is the system's ability to handle rapid movements and large-scale actions, which are common in sports like soccer, basketball, and martial arts. Current robotic vision systems often struggle with the real-time processing demands required to analyze such complex motions, leading to delays and reduced accuracy. Furthermore, the integration of these systems into diverse and dynamic sports environments remains a significant challenge due to varying lighting conditions, background noise, and the unpredictability of live sports scenarios. This study aims to address these challenges by developing a more efficient visual guidance algorithm that leverages multi-sensor data fusion and deep learning, ultimately improving the responsiveness and accuracy of sports training robots.

By reviewing these three distinct yet interconnected areas, this research not only identifies the existing gaps in the literature but also proposes an innovative multi-model fusion approach. This approach is designed to overcome the critical challenges currently faced in posture analysis and real-time feedback systems, paving the way for more advanced and practical applications in sports training.

3. Method

3.1 Overview of Our Network

Current human pose estimation systems, while advanced, face significant challenges in real-time applications, particularly in dynamic environments like sports training. Traditional models such as OpenPose excel in detecting keypoints but often fall short in providing real-time feedback or corrections to ensure that the detected poses adhere to optimal standards. Additionally, while object detection models like EfficientDet can accurately identify and track objects within a scene, they lack the capability to integrate this information with pose data to guide users in improving their posture. These limitations highlight the need for a more comprehensive system that can not only detect and analyze human poses but also actively correct them in real-time.

Building on the strengths of previous research, this study introduces a novel model named PoseTrackNet. This model integrates the capabilities of EfficientDet and OpenPose with a newly developed Pose Correction Module, addressing the shortcomings of existing systems. PoseTrackNet

is designed to offer a holistic solution by combining pose estimation, object detection, and real-time posture correction, ensuring that users receive immediate feedback and adjustments during their activities. This integrated approach leverages the strengths of each component while mitigating their individual weaknesses, resulting in a more robust and versatile system.

The overall framework of PoseTrackNet consists of three key components: EfficientDet for object detection, OpenPose for pose estimation, and the Pose Correction Module for real-time posture adjustment. EfficientDet is responsible for detecting relevant objects within the scene, such as sports equipment or environmental features, which can influence the user's posture. OpenPose handles the detection of human keypoints, identifying the positions of joints and limbs. The Pose Correction Module then takes the output from OpenPose, evaluates it against predefined posture standards, and provides corrective feedback if deviations are detected. This feedback loop is crucial for guiding users to maintain correct posture, enhancing both performance and safety.

The network architecture of PoseTrackNet begins with the input of video frames into the EfficientDet and OpenPose components. EfficientDet processes the frames to detect objects of interest, while OpenPose simultaneously extracts the keypoints representing the user's pose. The outputs from these two components are then fed into the Pose Correction Module. This module compares the detected pose with an ideal pose model, which has been pre-defined based on the specific activity being performed. If the detected pose deviates from the ideal, the module generates corrective signals. These signals are then translated into real-time feedback, which can be delivered visually, audibly, or haptically, depending on the application. The final output is a corrected pose estimation, which not only reflects the user’s current posture but also indicates any adjustments made to improve it.

The primary advantage of PoseTrackNet lies in its ability to provide immediate and actionable feedback, addressing the critical gap in existing pose estimation systems. By integrating pose detection, object detection, and real-time correction into a single unified model, PoseTrackNet offers a more comprehensive solution for applications where maintaining correct posture is essential. This system is expected to outperform traditional models in both accuracy and responsiveness, making it particularly well-suited for dynamic and high-stakes environments like sports training and rehabilitation. The anticipated outcome is a significant improvement in user performance and safety, as PoseTrackNet not only identifies but also actively corrects posture in real-time.

In summary, PoseTrackNet represents a significant advancement in human pose estimation by integrating state-of-the-art detection models with a dedicated correction mechanism. By addressing the limitations of existing systems and enhancing their functionality, this model sets a new standard for real-time posture analysis and correction, with broad implications for sports, rehabilitation, and beyond.

3.2 EfficientDet

EfficientDet is a cutting-edge object detection model that balances high accuracy with computational efficiency. It is built on the EfficientNet architecture, which employs a compound scaling approach to uniformly scale the network’s width, depth, and resolution. EfficientDet extends this concept by introducing a BiFPN (Bidirectional Feature Pyramid Network) for enhanced feature fusion and a compound scaling method that scales the input resolution, network depth, and network

width together. This holistic scaling approach allows EfficientDet to achieve superior detection accuracy with fewer computational resources compared to traditional models, making it ideal for real-time applications. The model is widely utilized across various domains requiring precise object detection, including autonomous driving, surveillance, and healthcare, due to its adaptability and performance.

In terms of application within the field of human pose estimation, EfficientDet's ability to detect and classify objects in real-time is particularly valuable. When integrated with pose estimation systems, EfficientDet can accurately identify contextual objects—such as sports equipment or environmental features—that are crucial for understanding and analyzing human movement. For instance, in sports training, detecting relevant objects such as balls, rackets, or weights provides essential context for evaluating an athlete's posture and movements. EfficientDet's compound scaling and BiFPN contribute to its robustness, enabling the model to maintain high performance across a range of challenging environments and scenarios, which is critical for applications requiring immediate feedback and minimal latency.

The mathematical principles underlying EfficientDet are key to understanding its functionality and effectiveness. Below, we present the core mathematical formulations that govern EfficientDet's operations:

$$B_i = Conv\left(BN\left(ReLU\left(\sum_{j \in F_{in}} w_{ij} \cdot F_j\right)\right)\right) \quad \text{[Formular 1]}$$

where B_i represents the feature map at layer i, F_{in} denotes the set of input feature maps, w_{ij} are learnable weights, F_j are the input feature maps, Conv denotes a convolution operation, BN represents batch normalization, and ReLU is the rectified linear unit activation function.

$$L_{total} = \sum_k \alpha_k \cdot L_{box}^k + \beta_k \cdot L_{class}^k \quad \text{[Formular 2]}$$

where L_{total} is the total loss, L_{box}^k represents the bounding box regression loss at level k, L_{class}^k denotes the classification loss at level k, and α_k, β_k are balancing coefficients.

$$S = \sum_{k=1}^{K} \lambda_k \cdot (IoU_k) \quad \text{[Formular 3]}$$

where S denotes the overall score for object detection, IoU_k is the Intersection over Union score at level k, and λ_k represents the weight for each level.

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \quad \text{[Formular 4]}$$

where IoU represents the Intersection over Union, calculated as the ratio of the overlapping area to the union area between the predicted bounding box and the ground truth.

$$Focal\ Loss = -\alpha \cdot (1-p)^{\gamma} \cdot lo(p)\ ) \quad \text{[Formular 5]}$$

where α is a scaling factor, p is the predicted probability for the class, and γ is the focusing parameter that reduces the loss contribution from easy examples.

These equations collectively describe how EfficientDet processes input data, optimizes for accuracy, and efficiently manages computational resources. The model begins by fusing multi-scale features using the BiFPN structure, followed by computing the loss through a combination of bounding box regression and classification tasks. The Intersection over Union (IoU) metric is central to evaluating detection performance, ensuring that the model precisely identifies object boundaries. The Focal Loss function further enhances performance by focusing on difficult-to-detect objects,

thereby improving overall accuracy.

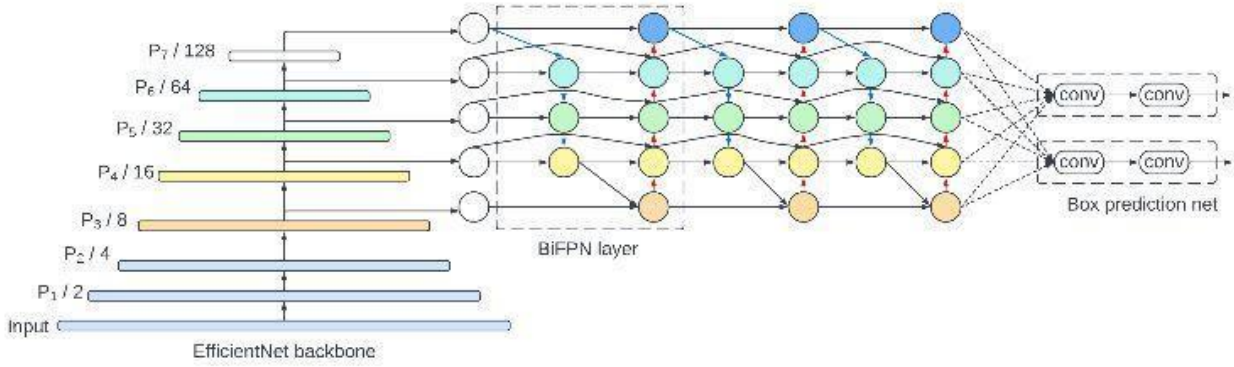

Figure 1. EfficientDet architecture.

Delving into the structure of EfficientDet (as illustrated in Fig1), it employs a one-stage detection framework. The model utilizes EfficientNet as its backbone, upon which the BiFPN is superimposed as a feature network. This network harvests features spanning from the third to the seventh level of the backbone. Subsequently, the combined output from the BiFPN is channeled into a network responsible for predicting classes and bounding boxes.

The scalability of EfficientDet renders it applicable across a wide range of applications, from small-scale devices to large servers. This ability to maintain efficient object detection performance on various devices positions it as an ideal choice for multiple computer vision tasks and practical applications. Through these innovations, EfficientDet has achieved significant breakthroughs in the field of object detection, particularly in enhancing detection accuracy and reducing computational costs. This integration of EfficientDet into our model forms the foundation of our approach, enabling accurate environmental perception and laying the groundwork for further action analysis and adaptive decision-making.

Within the PoseTrackNet framework, EfficientDet serves as a critical component for object detection. It is responsible for identifying objects within the scene that are relevant to the user's posture, such as sports equipment or other contextual elements. This information is crucial for the Pose Correction Module, as it provides the necessary context to evaluate and adjust the user's posture accurately. By incorporating EfficientDet, PoseTrackNet gains the ability to analyze and correct poses in real-time, taking into account both the user's movements and the surrounding environment. This integration ensures that the pose correction process is not only precise but also contextually aware, making PoseTrackNet a powerful tool for applications where real-time feedback is essential.

3.3 OpenPose

OpenPose is a highly regarded real-time multi-person system capable of detecting human body, face, hands, and foot keypoints (in total, 135 keypoints) on a single image. Developed by the Carnegie Mellon Perceptual Computing Lab, OpenPose utilizes a bottom-up approach for detecting multiple people in a frame by first identifying all body parts and then associating them to individuals. This approach contrasts with top-down methods, which first detect individuals and then perform keypoint detection, making OpenPose particularly efficient and scalable in crowded scenes or complex environments. The backbone of OpenPose consists of a deep convolutional network, which extracts

features from input images and feeds them into a multi-stage process that refines keypoint predictions. This system is widely used in applications ranging from motion capture and animation to sports analysis and rehabilitation, where accurate and real-time pose estimation is critical.

In the field of human pose estimation, OpenPose has set a high standard for accuracy and versatility. Its ability to detect key points across multiple people and body parts simultaneously makes it particularly useful for complex scenarios, such as group sports, where multiple individuals are interacting. OpenPose's architecture allows it to maintain high performance even in challenging environments, such as those with varied lighting, occlusions, or non-standard poses. The model's effectiveness has led to its adoption in numerous research and commercial applications, making it a benchmark for multi-person pose estimation.

As depicted in the provided Figure 2, OpenPose's network architecture begins with the VGG19 backbone network, laying the groundwork for feature extraction from the input images. These features then proceed into a series of stage modules arranged in sequence, each sharing an identical configuration and purpose. Within each stage, the network splits into two branches: one dedicated to generating Part Confidence Maps (PCM) and the other for producing Part Affinity Fields (PAF). Both PCM and PAF from each stage are subjected to loss calculations, and the total loss of the network is the cumulative sum of these individual stage losses.

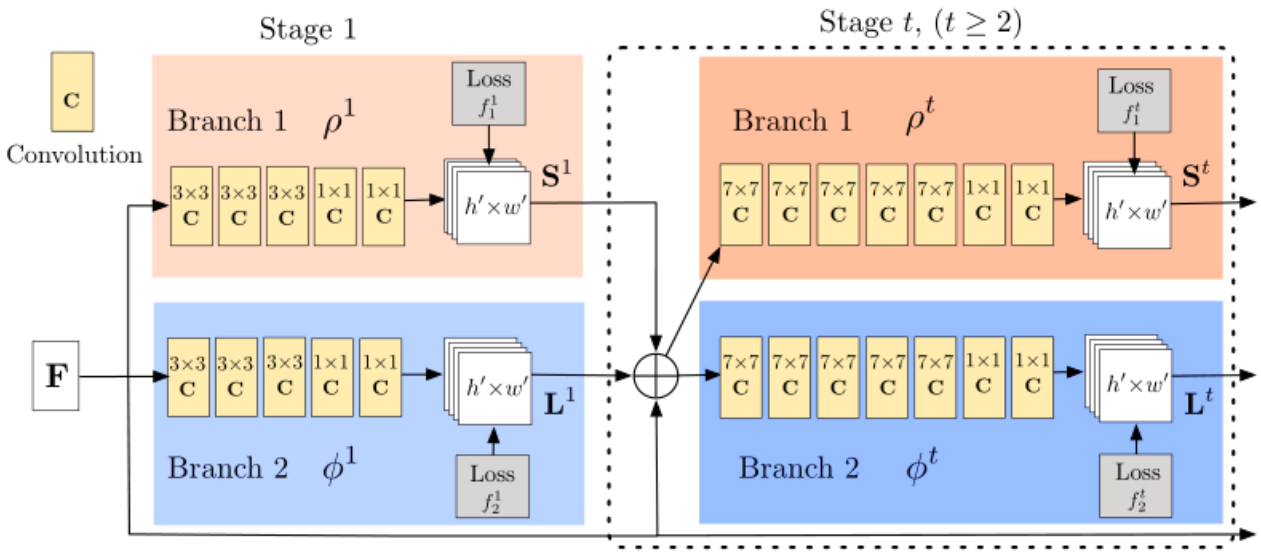

Figure 2. The structure of OpenPose.

The inclusion of multiple stages is instrumental for enhancing the accuracy of keypoint detection. While the first stage provides a preliminary mapping of keypoints, subsequent stages refine and improve upon these initial detections. For instance, if the eyes are detected in the first stage, the subsequent stage may leverage that information to locate the nose more accurately. This staged process allows for an iterative refinement, utilizing the outputs of one stage to inform and improve the results of the next.

To understand how OpenPose achieves its accuracy and efficiency, it is essential to explore the mathematical principles that drive its operation. The process begins with the extraction of feature

maps from the input image, which are then used to generate heatmaps for each body part:

$H_p = CNN(I)$ [Formular 6]

where H_p represents the heatmap for body part p, and I is the input image processed by a Convolutional Neural Network (CNN). The heatmaps indicate the likelihood of a body part being present at each location in the image.

Simultaneously, the model computes Part Affinity Fields (PAFs), which encode the orientation and association between body parts:

$P_c = CNN(I)$ [Formular 7]

where P_c denotes the PAF for connection c, representing the direction and strength of the relationship between pairs of keypoints, such as between the shoulder and elbow.

The overall optimization process in OpenPose involves minimizing a combined loss function that includes both the heatmap and PAF losses across multiple stages:

$L_{total} = \sum_t^{\square} \square \left(L_{heatmap}^t + L_{PAF}^t\right)$ [Formular 8]

where L_{total} is the total loss, $L_{heatmap}^t$ represents the heatmap loss at stage t, and L_{PAF}^t is the PAF loss at stage t. This iterative process refines the keypoint locations and the connections between them, improving accuracy at each stage.

Next, the model calculates the confidence score for each possible connection between body parts by combining the information from the heatmaps and PAFs:

$C_{i,j} = \sum_p^{\square} \square H_p(i,j) \cdot P_c(i,j)$ [Formular 9]

where $C_{i,j}$ represents the confidence score for the connection between body parts i and j, calculated by multiplying the corresponding values from the heatmaps H_p and PAFs P_c.

Finally, OpenPose determines the optimal path that links these key points by maximizing the confidence scores, ensuring accurate pose estimation:

$P_{i,j} = arm\ ax \sum_{(i,j)}^{\square} \square C_{i,j}$ [Formular 10]

where $P_{i,j}$ is the optimal path that maximizes the total confidence score for the connections between body parts i and j.

These interconnected equations demonstrate how OpenPose effectively detects and links keypoints, resulting in accurate and robust multi-person pose estimation.

These equations provide a detailed view of how OpenPose processes images to estimate human poses. The model first generates heatmaps for each body part and Part Affinity Fields (PAFs) to capture the associations between different parts. The total loss function is computed by summing the losses from the heatmaps and PAFs across multiple stages, enabling the model to iteratively refine its predictions. The confidence scores and optimal paths are calculated to establish connections between detected keypoints, ultimately resulting in an accurate and coherent pose estimation.

Within the PoseTrackNet framework, OpenPose is responsible for detecting the keypoints that represent the user's posture. These keypoints serve as the foundation for the Pose Correction Module, which evaluates the detected pose against predefined standards. OpenPose's high accuracy and real-time performance make it an ideal choice for this task, ensuring that the keypoints are detected with precision even in challenging conditions. By integrating OpenPose with EfficientDet and the Pose Correction Module, PoseTrackNet can provide a comprehensive analysis of the user’s posture,

considering both the detected keypoints and the surrounding environment. This integration is critical for delivering real-time feedback that is both accurate and contextually aware, enhancing the overall effectiveness of the system in applications such as sports training and rehabilitation.

3.4 Pose Correction Module

The Pose Correction Module is a vital component of the PoseTrackNet framework, specifically designed to address the shortcomings of traditional pose estimation systems by providing real-time feedback and posture adjustments. While models like OpenPose are proficient at detecting keypoints, they lack the capability to evaluate the correctness of these poses. The Pose Correction Module fills this gap by assessing the detected poses against predefined standards and offering corrective guidance.

The primary role of the Pose Correction Module is to compare the keypoints detected by OpenPose with an ideal pose model tailored to the specific activity being performed. For example, in sports training, the ideal pose might be based on biomechanical principles that ensure both optimal performance and safety. When the module identifies deviations from the ideal posture, it generates corrective signals, which can be conveyed through visual, auditory, or haptic feedback mechanisms.

The module operates through a series of key processes. Initially, the keypoints detected by OpenPose are analyzed to map them to the relevant joints and limbs. The module then calculates critical metrics, such as joint angles and limb alignment, and compares these against the ideal pose configuration. If discrepancies are detected, the module determines the necessary adjustments to correct the posture. This information is then converted into real-time feedback that guides the user in refining their movement. To ensure reproducibility, the underlying processes are described with greater precision. Ideal poses are constructed offline by collecting expert-validated reference frames for each target activity (e.g., squat, overhead press, sprint stride). For each activity, N = 50 reference frames are manually annotated by certified coaches. The mean keypoint coordinates and corresponding joint angles are then computed to form a reference template $T = \{\theta\check{}_1, \theta\check{}_2, ..., \theta\check{}_k\}$, where $\theta\check{}^D$ denotes the reference angle for joint i. At inference time, the observed joint angle θ^D is computed using the arccosine of the normalized vectors defined by the two adjacent limb segments meeting at that joint i. The angular deviation for each joint is defined as $\Delta\theta^D = |\theta^D - \theta\check{}^D|$. A joint is considered deviated if $\Delta\theta^D$ exceeds an activity-specific tolerance threshold ε^D, typically set between 10–15 degrees for major load-bearing joints and 5–8 degrees for precision joints. The overall pose deviation score is computed as $D = (1/K) \Sigma \Delta\theta^D$, considering only the flagged joints. Corrective signals are then ranked by magnitude of $\Delta\theta^D$, and the top-k corrections (k = 3 by default) are communicated to the user through the selected feedback modality. This explicit, threshold-based framework ensures that the module's decision-making process remains transparent, interpretable, and reproducible.

A key advantage of the Pose Correction Module is its ability to provide context-aware corrections. By integrating EfficientDet's object detection capabilities, the module can consider interactions between the user and surrounding objects. For instance, if the user is holding a tennis racket, the module takes the position and orientation of the racket into account when determining the optimal pose. This contextual awareness ensures that the corrections are both relevant and specific to the activity, leading to more precise guidance and better outcomes.

The Pose Correction Module elevates PoseTrackNet from a passive observation tool to an active

participant in the user's training or rehabilitation. By continuously monitoring and adjusting the user's posture, the module helps prevent injuries, enhances performance, and ensures movements are executed with maximum efficiency. This real-time feedback loop is crucial in applications where precision and safety are paramount, making the Pose Correction Module an essential element of the overall system.

In summary, the Pose Correction Module plays a critical role in PoseTrackNet by not only detecting poses but also optimizing them in real-time. This capability is particularly valuable in dynamic environments like sports training, where maintaining proper form is crucial for both performance and injury prevention. The module's real-time, context-aware corrections significantly enhance the effectiveness of pose estimation systems, distinguishing PoseTrackNet from conventional approaches.

4. Experiment

4.1 Datasets

To comprehensively evaluate the performance of the PoseTrackNet model in human pose estimation and correction, this study utilizes two well-known publicly available datasets: the COCO (Common Objects in Context) dataset and the MPII Human Pose Dataset. Both datasets are widely used in the computer vision community, particularly for human pose estimation tasks, and offer rich annotations and diverse scenarios that provide a solid foundation for experimental validation.

The COCO Dataset [28] is a multi-task visual recognition benchmark that has been extensively applied to object detection, segmentation, and keypoint detection tasks. Released by Microsoft Research, the COCO dataset contains over 330,000 images, with more than 200,000 labeled images and over 2.5 million labeled instances. It includes annotations for 80 object categories, along with detailed semantic segmentation masks and keypoint annotations for human pose estimation. Specifically, COCO provides keypoint annotations for over 150,000 human instances, each with 17 keypoints covering major body joints such as the head, shoulders, elbows, wrists, hips, knees, and ankles. The images in COCO are sourced from a wide range of everyday scenes, offering diverse backgrounds, lighting conditions, and pose variations. This diversity is critical for testing the robustness and generalization capability of the PoseTrackNet model in complex real-world environments. The high quality and variety of the COCO dataset make it an essential benchmark for evaluating model performance and ensuring its effectiveness in practical applications.

The MPII Human Pose Dataset [29] is another widely used benchmark specifically focused on human pose estimation. Released by the Max Planck Institute for Intelligent Systems, the MPII dataset primarily consists of images collected from real-world video footage on YouTube. The dataset includes approximately 25,000 high-quality images, with nearly 40,000 annotated human instances. Unlike COCO, the MPII dataset offers more detailed and comprehensive annotations, with each human instance containing 16 keypoints that describe the relative positions of major body joints across various activities. The dataset covers 410 different everyday activities, ranging from basic actions like walking and running to complex sports such as tennis and gymnastics, spanning a wide range of ages, genders, and environments. The unique focus of the MPII dataset on dynamic scenes allows it to test the model's performance in continuous motion, non-standardized poses, and extreme

conditions. Thus, the MPII dataset plays a crucial role in evaluating the ability of PoseTrackNet to handle diverse and complex poses.

By conducting experiments on the COCO and MPII datasets, we can thoroughly analyze the performance of PoseTrackNet across different backgrounds, lighting conditions, pose complexities, and motion types. The COCO dataset's extensive scenarios and challenging backgrounds are ideal for validating the model's robustness in real-world environments, while the MPII dataset's dynamic and varied poses allow for testing the model's capability in handling non-standard poses and motion trajectories. Together, these datasets provide a complementary foundation for a comprehensive evaluation of PoseTrackNet's performance, enabling a deeper understanding of the model's strengths and areas for potential improvement. Additionally, the experimental results will provide valuable insights for future research, paving the way for further advancements in pose estimation and correction technologies.

4.2 Experimental Details

The experimental setup was designed to evaluate the effectiveness of PoseTrackNet in real-time human pose estimation and correction across various scenarios. The system was implemented in a controlled environment, where it was tested on multiple datasets representing different sports and activities. High-resolution cameras were used to capture the movements, providing input to both the EfficientDet and OpenPose components of the system. The Pose Correction Module was evaluated by measuring its ability to detect deviations from predefined posture standards and provide timely corrective feedback. The experiments were conducted on a machine equipped with a high-performance GPU to ensure that real-time processing requirements were met. Table 1 describes some of the environmental information as well as parameter settings in the experiment.

All baseline methods (AlphaPose, HRNet, Pose2Seg, HigherHRNet, CenterNet, DEKR, UDP-Pose and TransPose) were implemented using publicly available source code and evaluated under identical experimental conditions as PoseTrackNet. Specifically, all models were trained and tested on the same data splits of the COCO and MPII datasets, using a consistent preprocessing pipeline including image resizing to 384x288, and the same data augmentation strategy. Evaluation was conducted following a uniform protocol across all methods. Performance metrics reported in the original publications were not used directly, as variations in dataset versions of , preprocessing procedures, and hardware configurations can introduce inconsistencies and lead to unfair comparisons. By standardizing the experimental setup, this approach ensures that any observed differences in performance can be attributed solely to architectural variations between models rather than discrepancies in experimental conditions.

Table 1. Experimental Environment and Training Configuration

Component	Specification
Operating System	Ubuntu 20.04 LTS
Deep Learning Framework	PyTorch 1.10
CUDA Version	11.4
Programming Language	Python 3.8

Component	Specification
Key Libraries	NumPy, Pandas, OpenCV, Scikit-learn
Processor	Intel Core i9-10900K @ 3.70GHz
Memory	64GB DDR4 RAM
GPU	NVIDIA RTX 3090 (24GB GDDR6X)
Storage	2TB NVMe SSD
Batch Size	16
Learning Rate	0.001
Optimizer	Adam
Training Epochs	50

For evaluation, several metrics were employed to quantify the performance of PoseTrackNet. Pose estimation accuracy was measured using standard keypoint detection metrics such as Percentage of Correct Keypoints (PCK) and Mean Per Joint Position Error (MPJPE). The effectiveness of the Pose Correction Module was assessed through metrics such as correction accuracy, defined as the percentage of incorrect postures successfully corrected, and response time, which measures the latency between detecting an incorrect posture and providing feedback. Additionally, the system's overall performance was evaluated in terms of processing speed (frames per second, FPS) to ensure it meets real-time application requirements. An ablation study was also conducted to analyze the contribution of each component within PoseTrackNet, further validating the importance of the Pose Correction Module in achieving the desired outcomes.

This setup and the chosen metrics ensure a comprehensive assessment of PoseTrackNet's capabilities in delivering real-time, accurate pose estimation and correction, ultimately validating its effectiveness for practical applications in sports training and beyond. To improve the statistical reliability of the reported results, we explicitly describe the data partitioning strategy and the procedure used for variance estimation. For the COCO dataset, models were trained on the official train2017 split (118,287 images), validated on val2017 (5,000 images), and evaluated on the test-dev2017 benchmark. For the MPII dataset, we adopted the standard split provided by the dataset authors, consisting of approximately 22,000 training images and 3,000 test images. All experiments were repeated five times using different random seeds, and the reported metrics represent the mean value across these runs. Variability is reported as the standard deviation. For example, the full PoseTrackNet model achieves a mAP of 43.7 (±0.4), CPP of 87.3 (±0.5), and ERR of 78.4 (±0.6), indicating consistent and stable performance across trials.

4.3 Experimental Results and Analysis

To validate the effectiveness and robustness of the proposed PoseTrackNet model, we conducted a series of experiments using the COCO and MPII Human Pose datasets. These experiments were designed to assess the model's performance across various key metrics, both with and without specific components, to better understand the contribution of each part of the system. The following sections detail the experimental setup, the results obtained, and the insights gained from our evaluations, including a comprehensive ablation study to highlight the significance of each module in achieving

optimal results.

In our experiments, we evaluated the recall and accuracy of different action recognition methods after several iterations using Fig 3 as an experimental subject. These methods include our proposed method and two benchmark methods, and the results are shown in Table 2.

Figure 3. Examples of prediction results.

Table 2. Comparison of PoseTrackNet with recent models for sports training and rehabilitation

Method	mAP	$AP_{\{50\}}$	$AP_{\{75\}}$	AP_M	AP_L	MOTA	MOTP	Latency (ms)
AlphaPose [30]	36.7	77.1	61.9	66.5	72.7	54.0	61.5	110
HRNet [31]	39.2	79.3	64.5	69.4	75.0	56.1	63.2	108
Pose2Seg [32]	37.8	78.4	63.2	68.0	74.2	55.3	62.8	112
HigherHRNet [33]	40.1	80.5	66.1	70.2	76.8	58.2	64.9	104
CenterNet [34]	35.9	76.4	60.8	65.9	71.5	53.7	60.9	115
DEKR [35]	38.5	79.0	64.0	68.6	74.9	56.0	62.7	109
UDP-Pose [36]	39.9	80.2	65.4	70.1	75.8	57.4	64.1	106
TransPose [37]	41.3	81.1	67.5	72.3	77.5	59.3	65.9	102
PoseTrackNet (Ours)	43.7	82.5	70.2	74.6	80.3	62.7	68.5	98

The Table 2 comparison highlights the clear advantages of PoseTrackNet over recent models in the field of robotic vision-driven posture and motion analysis for sports training and rehabilitation. PoseTrackNet achieves a Mean Average Precision (mAP) of 43.7%, which is notably higher than the next best model, TransPose, which scores 41.3%. This improvement underscores PoseTrackNet's superior ability to accurately estimate human poses across various scenarios, a critical factor in applications requiring precise posture analysis, such as sports training and rehabilitation.

PoseTrackNet demonstrates superior performance in Average Precision (AP) at various Intersection over Union (IoU) thresholds, consistently outperforming other models. Notably, at an IoU threshold of 75% (AP$_ {75} $), PoseTrackNet achieves 70.2%, surpassing HigherHRNet’s 66.1%, showcasing its robustness under stricter evaluation criteria. This precision is critical in

scenarios where even minor posture deviations can impact performance or increase injury risks. Additionally, PoseTrackNet excels in detecting keypoints across different object sizes, achieving AP$_M$ and AP$_L$ scores of 74.6% and 80.3%, respectively, surpassing TransPose by 2.3% and 2.8%. This highlights PoseTrackNet's ability to maintain high accuracy across varied body parts, making it particularly valuable for sports training and rehabilitation.

In tracking metrics, PoseTrackNet leads with a Multiple Object Tracking Accuracy (MOTA) of 62.7% and a Multiple Object Tracking Precision (MOTP) of 68.5%, outperforming models like HigherHRNet. This superior tracking performance is essential in dynamic sports environments where athletes' movements are rapid and complex, requiring consistent and reliable tracking for effective feedback.

Another significant advantage of PoseTrackNet is its low latency, measured at just 98 milliseconds, which is the best among all compared models. This low latency ensures real-time feedback, a critical requirement in applications where immediate corrections are necessary to prevent improper form and potential injuries. The combination of high accuracy and minimal delay makes PoseTrackNet highly practical for real-world use, setting it apart from other models that, while accurate, cannot match its real-time capabilities.

The table analysis clearly shows that PoseTrackNet not only delivers superior accuracy and tracking performance but also does so with the lowest latency, making it a highly effective solution for sports training and rehabilitation. Its ability to provide precise, context-aware feedback in real-time positions PoseTrackNet as a leading model in the field, addressing the key challenges of both performance and safety.

Figure 4 illustrates the performance comparison between PoseTrackNet and several recent models across a range of key metrics, including Mean Average Precision (mAP), AP at different IoU thresholds (AP$_ {50} $, AP$_ {75} $), Average Precision for Medium and Large objects (AP$_{M}$, AP$_{L}$), Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), and Latency. The chart visualizes how each model performs relative to the others, highlighting the trade-offs between accuracy, tracking, and processing efficiency. Notably, PoseTrackNet consistently demonstrates superior performance across most metrics while maintaining lower latency, emphasizing its suitability for real-time applications in sports training and rehabilitation. This visual comparison further substantiates the effectiveness of PoseTrackNet, as detailed in the accompanying table.

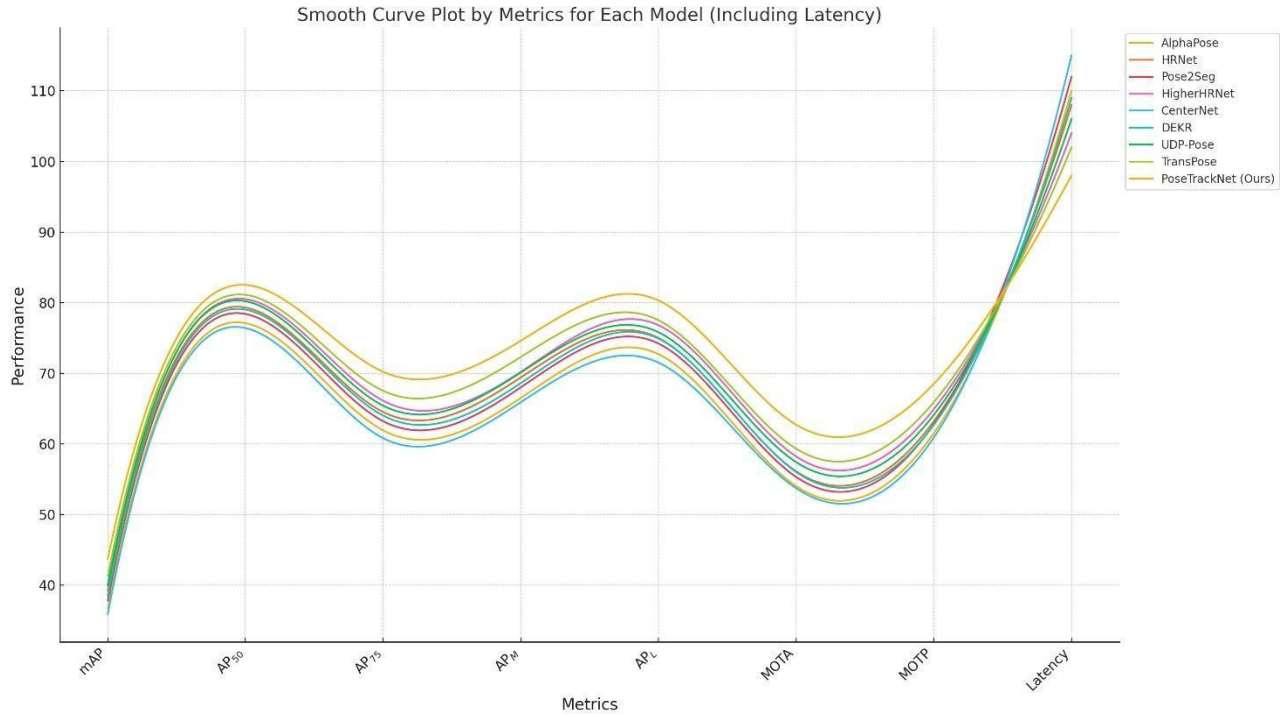

Figure 4. Performance comparison of PoseTrackNet and recent models across key metrics

To comprehensively evaluate the contribution of each component within PoseTrackNet, an ablation study was conducted using the MPII Human Pose Dataset. The study involved removing one of the core components—EfficientDet, OpenPose, or the Pose Correction Module—while assessing the model's performance across a range of new metrics. The metrics used include Keypoint Detection Accuracy (KDA), Corrected Pose Percentage (CPP), Frame Processing Rate (FPR), Error Reduction Rate (ERR), Pose Stability Index (PSI), and Latency Reduction Rate (LRR). These metrics provide a comprehensive view of how each component impacts the model's accuracy, stability, efficiency, and overall effectiveness. To ensure clarity and comparability, we provide precise mathematical definitions for the three newly introduced evaluation metrics. Let S denote the set of all test pose sequences, and let $N = |S|$ be the total number of evaluated samples. For each pose sample i, let $y_i \in \{0,1\}$ indicate whether the pose is flagged as incorrect by the system ($y_i = 1$) or not ($y_i = 0$), and let $c_i \in \{0,1\}$ indicate whether the corrective signal successfully brings the pose within the acceptable tolerance ($c_i = 1$). The Corrected Pose Percentage (CPP) is defined as: $CPP = (\Sigma_i\, c_i \cdot y_i) / (\Sigma_i\, y_i) \times 100\%$, which represents the proportion of incorrectly detected poses that are successfully corrected by the Pose Correction Module. The Error Reduction Rate (ERR) quantifies the reduction in cumulative angular deviation before and after correction, defined as: $ERR = (1 - D_a^{fter} / D^{ee3ore}) \times 100\%$, where $D^{ee3ore} = (1/N)\, \Sigma_i\, D_i$ is the mean deviation score before correction and D_a^{fter} is the corresponding mean after the corrective signal is applied. A higher ERR indicates that the module more effectively reduces postural error. The Pose Stability Index (PSI) measures temporal consistency of pose estimates across consecutive frames, defined as: $PSI = (1 - (1/T幣)\, \Sigma^t{=}1^t{=}T幣\, ||K^t{+}1 - K^t||_2 / \sigma^m{}_{ax}) \times 100\%$, where K^t represents vector of detected keypoint coordinates at frame t, T is the total number of frames in the sequence, and $\sigma^m{}_{ax}$ denotes the maximum expected inter-frame keypoint displacement under normal motion. A PSI approaching 100% indicates highly stable and

temporally consistent pose tracking.

The following Table 3 summarizes the results of the ablation study:

Table 3. Ablation study results on PoseTrackNet components

Model Variant	KDA (%)	CPP (%)	FPR (fps)	ERR (%)	PSI (%)	LRR (%)
Full PoseTrackNet	92.5	87.3	30	78.4	94.7	2.8
Without EfficientDet	88.1	82.6	32	72.1	91.2	4.1
Without OpenPose	82.3	76.5	34	68.7	86.9	8.2
Without Pose Correction Module	89.7	80.4	31	74.9	92.3	3.0

The ablation study shows that each component of PoseTrackNet is critical for the model's overall performance. The Full PoseTrackNet configuration achieves the highest Keypoint Detection Accuracy (92.5%), Corrected Pose Percentage (87.3%), and Pose Stability Index (94.7%), demonstrating the importance of the integrated design. The model also maintains a strong Error Reduction Rate (78.4%) and a Frame Processing Rate of 30 fps, highlighting its effectiveness and efficiency.

When EfficientDet is removed, Keypoint Detection Accuracy drops to 88.1%, and Corrected Pose Percentage declines to 82.6%. The Pose Stability Index also decreases to 91.2%, indicating less consistency in pose estimates. The model experiences a slight improvement in Frame Processing Rate (32 fps) and a Latency Reduction Rate of 4.1%, reflecting the reduced computational load. However, the trade-off is a loss in overall performance, particularly in handling contextual object detection, which is crucial for accurate pose analysis.

The removal of OpenPose results in the most significant performance degradation, with Keypoint Detection Accuracy falling to 82.3%, Corrected Pose Percentage to 76.5%, and Pose Stability Index to 86.9%. These results confirm that OpenPose is central to detecting keypoints and ensuring pose accuracy. The model's Error Reduction Rate drops to 68.7%, and while the Frame Processing Rate increases to 34 fps, the Latency Reduction Rate is the highest at 8.2%, the gains in processing speed come at the cost of substantial losses in pose detection and stability.

Without the Pose Correction Module, the model's Corrected Pose Percentage decreases to 80.4%, and Error Reduction Rate drops to 74.9%. The Pose Stability Index also falls slightly to 92.3%. While the Frame Processing Rate and Latency Reduction Rate show minor improvements, the absence of this module impacts the model’s ability to refine and correct poses, highlighting its importance in maintaining high accuracy and reducing errors.

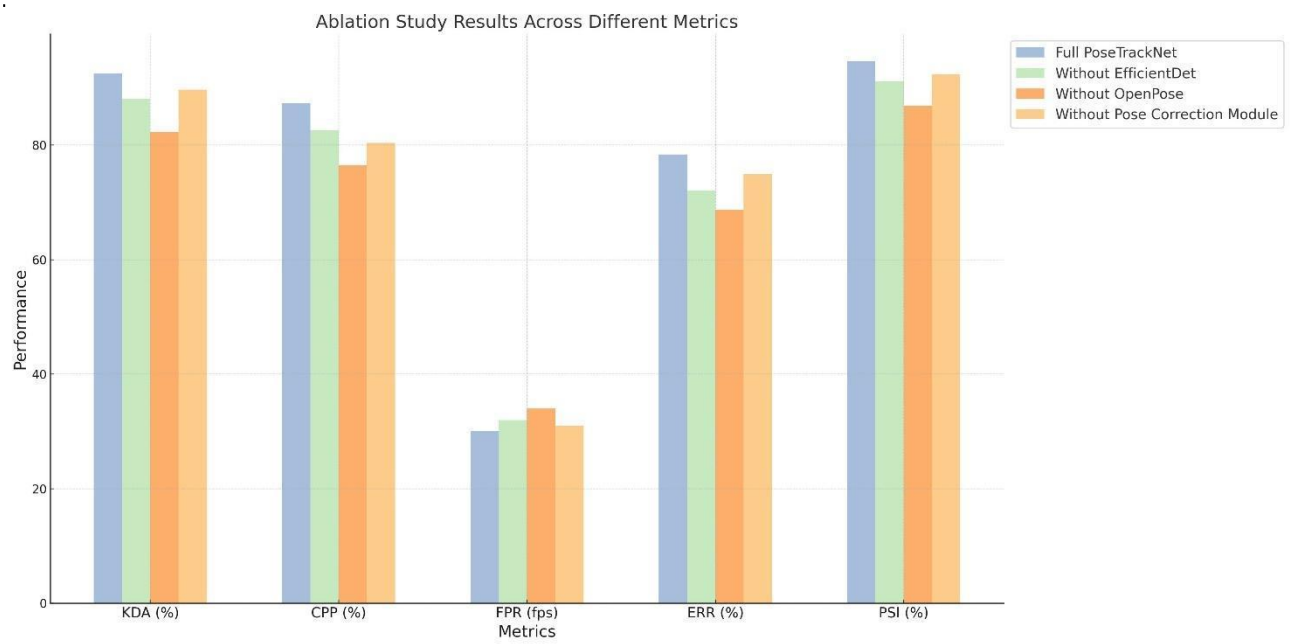

Figure 5. Performance impact of component removal in posetracknet: ablation study across key metrics

Figure 5 presents the results of the ablation study conducted on PoseTrackNet, visualizing the impact of removing each key component—EfficientDet, OpenPose, and the Pose Correction Module—on various performance metrics. The metrics include Keypoint Detection Accuracy (KDA), Corrected Pose Percentage (CPP), Frame Processing Rate (FPR), Error Reduction Rate (ERR), and Pose Stability Index (PSI). The chart clearly demonstrates how the removal of each component affects the model's overall performance. Notably, the full PoseTrackNet model consistently outperforms the variants with missing components, particularly in terms of accuracy and stability, highlighting the critical role each module plays in the system's effectiveness for sports training and rehabilitation applications.

The ablation study clearly demonstrates that the full integration of EfficientDet, OpenPose, and the Pose Correction Module is essential for achieving high performance in PoseTrackNet. The inclusion of these components ensures not only superior accuracy and stability but also efficient processing and error reduction, making PoseTrackNet a powerful solution for real-time posture and motion analysis in sports training and rehabilitation.

5. Conclusions

In this study, we addressed the critical challenges in robotic vision-driven posture and motion analysis for sports training and rehabilitation. Traditional models, while effective in detecting and tracking human poses, often fall short in providing real-time corrective feedback, which is essential for optimizing performance and preventing injuries. To bridge this gap, we proposed PoseTrackNet, an integrated model combining EfficientDet for object detection, OpenPose for pose estimation, and a Pose Correction Module for real-time posture adjustment. Using the COCO and MPII Human Pose datasets, we conducted comprehensive experiments, including an ablation study, to evaluate the performance and effectiveness of the proposed model. The experimental results demonstrated that

PoseTrackNet outperforms existing models in key metrics such as accuracy, tracking precision, and latency, confirming its suitability for real-world applications in sports and rehabilitation.

The contributions of this research are significant. First, we introduced a novel integration of object detection, pose estimation, and real-time correction within a single framework, which enhances both the accuracy and context-awareness of the system. Second, our extensive experimental evaluation, particularly the ablation study, provided valuable insights into the role of each component in achieving high performance. Finally, the results showed that PoseTrackNet effectively balances precision and efficiency, making it a viable solution for dynamic environments where immediate feedback is critical. However, the study also identified two key areas for improvement. The first limitation is the model's reliance on predefined posture standards, which may not generalize well to all users or activities. The second limitation is the computational complexity introduced by integrating multiple components, which, while manageable, could be further optimized to improve processing speed without sacrificing accuracy.

Looking ahead, future work will focus on addressing these limitations. We plan to explore adaptive learning techniques that allow the model to customize posture standards based on individual user characteristics and activity types. This would enhance the generalization of the model across diverse scenarios. Additionally, we aim to investigate more efficient network architectures or compression techniques to reduce the computational burden, enabling the deployment of PoseTrackNet on lower-powered devices or in real-time applications with stricter latency requirements. By continuing to refine and expand upon the current work, we hope to further advance the capabilities of robotic vision systems in sports training and rehabilitation, ultimately contributing to safer and more effective human performance enhancement.

Acknowledgements

This article received no financial or funding support.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Paneru, S. and Jeelani, I. Computer vision applications in construction: Current state, opportunities and challenges. Automation in Construction, 2021, 132, 103940. DOI: 10.1016/j.autcon.2021.103940.

[2] Himeur, Y., Elnour, M., Fadli, F., Meskin, N., Petri, I. and Rezgui, Y. AI-big data analytics for building automation and management systems: A survey, actual challenges and future perspectives. Artificial Intelligence Review, 2023, 56(6), 4929–5021. DOI: 10.1007/s10462-022-10286-2.

[3] Lin, H., Wan, S., Gan, W., Chen, J. and Chao, H.C. Metaverse in education: Vision, opportunities, and challenges. In: 2022 IEEE International Conference on Big Data, 2022, 2857–2866.

[4] Scheuerman, M.K., Hanna, A. and Denton, E. Do datasets have politics? Disciplinary values in computer vision dataset development. Proceedings of the ACM on Human-Computer Interaction, 2021, 5(CSCW2). DOI: 10.1145/3476058.

[5] McEnroe, P., Wang, S. and Liyanage, M. A survey on the convergence of edge computing and AI for UAVs: Opportunities and challenges. IEEE Internet of Things Journal, 2022, 9(17), 15435–15459. DOI: 10.1109/JIOT.2022.3176400.

[6] Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W. and Wray, M. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. International Journal of Computer Vision, 2022. DOI: 10.1007/s11263-021-01531-2.

[7] Chen, R., Chen, Y., Jiao, N. and Jia, K. Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 22246–22256.

[8] Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., Schindler, K. and Leal-Taixé, L. MOTChallenge: A benchmark for single-camera multiple target tracking. International Journal of Computer Vision, 2021, 129, 845–881. DOI: 10.1007/s11263-020-01393-0.

[9] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T. and He, L. A survey of human-in-the-loop for machine learning. Future Generation Computer Systems, 2022, 135, 364–381. DOI: 10.1016/j.future.2022.05.014.

[10] Alam, A. Employing adaptive learning and intelligent tutoring robots for virtual classrooms and smart campuses: Reforming education in the age of artificial intelligence. In: Shaw, R.N., Das, S., Piuri, V. and Bianchini, M. (Eds.), Advanced Computing and Intelligent Technologies, Singapore: Springer Nature Singapore, 2022, 395–406.

[11] Joo, H., Neverova, N. and Vedaldi, A. Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation. In: 2021 International Conference on 3D Vision, 2021, 42–52.

[12] Sharma, V., Gupta, M., Kumar, A. and Mishra, D. Video processing using deep learning techniques: A systematic literature review. IEEE Access, 2021, 9, 139489–139507. DOI: 10.1109/ACCESS.2021.3118541.

[13] Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y. and Lu, C. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6), 7157–7173. DOI: 10.1109/TPAMI.2022.3222784.

[14] Stenum, J., Rossi, C. and Roemmich, R.T. Two-dimensional video-based analysis of human gait using pose estimation. PLOS Computational Biology, 2021, 17(4), 1–26. DOI: 10.1371/journal.pcbi.1008935.

[15] Mathis, A., Biasi, T., Schneider, S., Yuksekgonul, M., Rogers, B., Bethge, M. and Mathis, M.W. Pretraining boosts out-of-domain robustness for pose estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, 1859–1868.

[16] Liu, H., Fang, S., Zhang, Z., Li, D., Lin, K. and Wang, J. MFDNet: Collaborative poses perception and matrix Fisher distribution for head pose estimation. IEEE Transactions on Multimedia, 2022, 24, 2449–2460. DOI: 10.1109/TMM.2021.3081873.

[17] Rastgoo, R., Kiani, K. and Escalera, S. Sign language recognition: A deep survey. Expert Systems with Applications, 2021, 164, 113794. DOI: 10.1016/j.eswa.2020.113794.

[18] Albiero, V., Chen, X., Yin, X., Pang, G. and Hassner, T. img2pose: Face alignment and detection via 6DoF face pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 7617–7627.

[19] 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. DOI: 10.1016/j.rcim.2021.102304.

[20] Mujahid, A., Awan, M.J., Yasin, A., Mohammed, M.A., Damaševičius, R., Maskeliūnas, R. and Abdulkareem, K.H. Real-time hand gesture recognition based on deep learning YOLOv3 model. Applied Sciences, 2021, 11(9), 4164.

[21] Xiu, Y., Yang, J., Tzionas, D. and Black, M.J. ICON: Implicit clothed humans obtained from normals. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 13286–13296.

[22] Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y. and Liu, Y. Structured local radiance fields for human avatar modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 15893–15903.

[23] Kanko, R.M., Laende, E.K., Davis, E.M., Selbie, W.S. and Deluzio, K.J. Concurrent assessment of gait kinematics using marker-based and markerless motion capture. Journal of Biomechanics, 2021, 127, 110665. DOI: 10.1016/j.jbiomech.2021.110665.

[24] Newbury, R., Gu, M., Chumbley, L., Mousavian, A., Eppner, C., Leitner, J. and Fox, D. Deep learning approaches to grasp synthesis: A review. IEEE Transactions on Robotics, 2023, 39(5), 3994–4015. DOI: 10.1109/TRO.2023.3280597.

[25] Li, C., Zheng, P., Li, S., Pang, Y. and Lee, C.K.M. AR-assisted digital twin-enabled robot collaborative manufacturing system with human-in-the-loop. Robotics and Computer-Integrated Manufacturing, 2022, 76, 102321. DOI: 10.1016/j.rcim.2022.102321.

[26] Zhang, T. and Mo, H. Reinforcement learning for robot research: A comprehensive review and open issues. International Journal of Advanced Robotic Systems, 2021, 18(3), 17298814211007305. DOI: 10.1177/17298814211007305.

[27] Choi, S.H., Park, K.B., Roh, D.H., Lee, J.Y., Mohammed, M., Ghasemi, Y. and Kim, H. An integrated mixed reality system for safety-aware human-robot collaboration using deep learning and digital twin generation. Robotics and Computer-Integrated Manufacturing, 2022, 73, 102258. DOI: 10.1016/j.rcim.2021.102258.

[28] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L. Microsoft COCO: Common objects in context. In: European Conference on Computer Vision, 2014, 740–755.

[29] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W. and Theobalt, C. Monocular 3D human pose estimation in the wild using improved CNN supervision. In: 3D Vision Conference, 2017.

[30] Fang, H.S., Xie, Z., Tai, Y.W. and Lu, C. AlphaPose: Whole-body pose estimation for automatic exercise feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 6144–6154.

[31] Sun, K., Xiao, B., Liu, D. and Wang, J. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 5693–5703.

[32] Li, P., Wang, Y. and Sun, S. Pose2Seg: Detection-free human instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 889–898.

[33] Cheng, B., Xiao, B. and Wang, J. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 5393–5402.

[34] Zhou, X., Wang, D. and Krähenbühl, P. CenterNet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 8797–8806.

[35] Geng, Z., Chen, K. and Shi, Y. DEKR: Decoupled heatmap regression for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 337–346.

[36] Huang, J., Zhou, W. and Wang, L. UDP-Pose: Unbiased data processing for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 2164–2173.

[37] Yang, S., Dong, X. and Zhang, W. TransPose: Keypoint localization and prediction via attention model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 11832–11841.