HR-YOLOv8-DE: Advancing Human-Robot Interaction with Precision Pose Estimation in Sports Rehabilitation

Rajesh Kumar KV*

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

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

DOI: https://doi.org/10.30211/JIC.202503.019

Submitted: Oct. 24, 2025 Accepted: Jan. 13, 2026

ABSTRACT

The integration of human-robot interaction within sports rehabilitation marks a significant advancement in improving the precision and effectiveness of therapeutic exercises. However, existing models often fail to accurately capture and analyze the intricate, dynamic movements required in rehabilitation, resulting in inadequate feedback and less optimal patient outcomes. To address these challenges, we propose the HR-YOLOv8-DE network, which integrates Diverse Branch Block (DBB), HRNet, and Efficient Multi-Scale Attention (EMA). This model is designed to enhance multi-scale feature extraction, maintain high-resolution pose estimation, and adaptively prioritize relevant features across varying conditions. Experimental results demonstrate the superior performance of our approach, with the HR-YOLOv8-DE model achieving a PCK of 82.0% on the MPII dataset and a mAP of 74.7% on the COCO dataset, significantly outperforming existing methods. These advancements not only improve the accuracy and adaptability of human motion analysis in rehabilitation but also set a new standard for future developments in robotic-assisted therapeutic interventions.

Keywords: Human-robot interaction, Sports rehabilitation, Pose estimation, HRNet, Multi-scale feature extraction, Efficient multi-scale attention.

1. Introduction

Sports and exercise have long been recognized as integral components of a healthy lifestyle, and their incorporation into rehabilitation programs can provide holistic benefits to patients. As an essential rehabilitative approach, sports rehabilitation aims to help patients restore impaired motor functions, enhancing their quality of life and autonomy [1]. This mode of rehabilitation is not only applicable to individuals recovering from sports injuries but also extends to patients affected by neurological disorders, chronic pain, and other health issues [2]. While sports rehabilitation plays a vital role in fostering recovery and improving quality of life, it still faces numerous challenges in practice. In modern society, there is a growing recognition of the importance of sports rehabilitation, particularly with the increasing aging population and lifestyle changes. However, traditional rehabilitation methods often encounter limitations such as a lack of personalization, difficulties in

measuring outcomes, and challenges in maintaining long-term commitment [3]. Moreover, conventional rehabilitation programs typically require patients to visit medical facilities for training, posing inconvenience and discomfort for those with mobility issues. Therefore, enhancing the personalization, effectiveness, and sustainability of sports rehabilitation has become an urgent issue in the field. Amidst these challenges, technological advancements have introduced new opportunities for sports rehabilitation. Multimodal robots, as emerging rehabilitative aids, have garnered widespread attention [4]. These robots integrate various sensory modalities, such as vision, sound, and touch, and are equipped with a degree of intelligence and adaptability. By interacting with patients, multimodal robots can monitor their movement in real-time, provide personalized training and guidance, and offer an improved rehabilitation experience [5]. In this context, research into leveraging multimodal robots for more intelligent and effective human-machine interaction in sports rehabilitation has begun to flourish. This research encompasses not just the hardware design and technical realization of robots but also the application of artificial intelligence in human-machine interactions. AI technologies, particularly machine learning and deep learning, have made significant strides, offering robust support for the human-machine interaction capabilities of multimodal robots in sports rehabilitation [6, 7]. With the guidance of AI, multimodal robots are poised to become vital assistants in future sports rehabilitation, providing more personalized and effective services to patients, and promoting enhanced and accelerated recovery outcomes.

Extensive investigations into human-robot interaction (HRI) with multimodal robotic systems have been conducted in the field of sports rehabilitation, aiming to address the limitations of traditional rehabilitation methods via advanced technological approaches and deliver more intelligent and personalized rehabilitation services. One study is an HRI-enhanced multimodal robotic system of sports rehabilitation based on the integration of an attention mechanism to dynamically weight information of the various sensory modalities used to identify and analyze movement behaviors of patients although its design and implementation needs further optimization to make it applicable and stable in complex environmental settings. The other study suggested a multimodal robotic platform which uses a better transformer model at natural language understanding and generation, with visual and auditory perception, to ease HRI in sport rehabilitation scenarios. Although this system has proved much progress in language understanding, it remains weak in action identification and act direction in the real-life interaction situation [9]. Also, one of the developments is the use of reinforcement learning algorithms to create individual rehabilitation guidelines. The model in question uses a deep Q-network (DQN) to discover the best rehabilitation measures that can be applied to the requirements of each patient and their progress [10]. The system has brought significant improvements in rehabilitation outcomes through the simulation of different rehabilitation scenarios as well as optimization of action strategies through reinforcement learning. Nevertheless, the model's sample complexity and sensitivity to reward function design present challenges for real-world deployment, necessitating further exploration and refinement. While their system has made notable progress in personalized rehabilitation guidance, improvements are still needed in the model's interpretability and explainability. These studies provide various approaches and methods for human-robot interaction with multimodal robots in sports rehabilitation. However, each method has certain limitations, such as issues with real-time performance, stability, language understanding, action

recognition, and multimodal information fusion.

In an attempt to eliminate constraints of previous research, we trained the HR-YOLOv8-DE network of sports rehabilitation integrated with human-robot interaction. This network consists of 3 important additions. First, we have substituted the Bottleneck block at YOLOv8 with the Diverse Branch Block (DBB) which improves the feature extraction of the network. Second, we incorporated the High-Resolution Network (HRNet) to enhance the accuracy of the pose estimation and fine details. Lastly, we have incorporated the Efficient Multi-Scale Attention (EMA) mechanism to enhance the perception of multi-scale features to allow our model to give more precise motion tracking and customized advice to rehabilitation robots. This assists the patients to do more productive rehabilitation exercises. In addition, the integrated technologies make the model very robust and generalize well, thus being applicable to a wide range of rehabilitation scenarios and patient population. To sum up, the HR-YOLOv8-DE network is a novel method of HRI-integrated sports rehabilitation study, which has both theoretical and practical importance.

Here are the three main contributions of the paper:

- This paper introduces a novel HR-YOLOv8-DE network that integrates Diverse Branch Block (DBB), HRNet (High-Resolution Network), and Efficient Multi-Scale Attention (EMA) mechanisms. Compared to existing models, the HR-YOLOv8-DE network demonstrates significant improvements in feature extraction, multi-scale feature perception, and high-resolution pose estimation, thereby enhancing human-robot interaction in sports rehabilitation.
- By incorporating HRNet, the proposed model achieves higher accuracy in pose estimation, capturing and analyzing detailed patient movements more effectively. This enhancement allows the robotic system to provide more precise and personalized rehabilitation guidance, significantly improving patient recovery outcomes.
- The integration of the EMA mechanism enables the HR-YOLOv8-DE network to adaptively prioritize relevant features under varying conditions, enhancing its ability to perceive multi-scale features. This advancement improves the model's robustness and generalization capabilities in complex environments, making it more effective across diverse rehabilitation settings and patient populations.

These contributions not only advance the field of robotic-assisted sports rehabilitation but also establish new benchmarks for future developments in rehabilitation technology.

2. Literature Review

2.1 Advancements in Multimodal Data Fusion and Motion Analysis for Rehabilitation Robotic

Recent advancements in the field of rehabilitation robotics have increasingly focused on the integration of multimodal data fusion techniques and real-time motion analysis technologies to enhance patient outcomes. Multimodal data fusion techniques aim to improve human-computer interaction by integrating data from various perceptual modalities, such as vision, speech, and touch, to provide a comprehensive understanding of patients' movement states. These methods involve designing and optimizing models that fuse different data modalities to enhance the monitoring and analysis of patient behaviors, thereby improving the accuracy and responsiveness of rehabilitation systems [11, 12]. Despite progress, challenges such as suboptimal data fusion results and low

accuracy in recognizing complex movements remain, necessitating further research to improve model architectures and fusion strategies.

Real-time motion data collection and processing technologies are essential in rehabilitation that records and measures patient motions. Such technologies include the implementation and integration of different sensors, development of data acquisition systems to guarantee proper data transmission and storage [13, 14]. Real-time data processing consists of such steps as data preprocessing, feature extraction, action recognition, and motion trajectory analysis (which is a strong background support of real-time monitoring and feedback) [15, 16]. The current problems are optimization of the collection of various data streams with synchronization and improved stability and accuracy of the system when operating in a complex environment. The next step of the research is to utilize the latest data mining and machine learning to enhance the real-time motion data technologies and become more intelligent and custom-made rehabilitation robot systems.

2.2 Application of Detection, Tracking, and Reinforcement Learning Techniques in Rehabilitation Robotics

The combination of detection and tracking systems and reinforcement learning algorithms is a new area of research in the intelligent rehabilitation robotics field. The target detection and tracking technologies are developed to provide the ability to perform real-time monitoring and guidance by identifying and tracking the movement targets of patients (body parts or rehabilitation equipment) in dynamic environments [17, 18]. These technologies provide essential data regarding the movement patterns of patients, which can be further used to identify the appropriate action and track the current trajectory. Nonetheless, issues exist in simultaneous sensing and tracking of a number of moving targets under different lighting conditions and body postures that undermines the stability and strength of the system [19]. Further studies are needed to come up with better detection and tracking algorithms, combining deep learning and conventional computer vision methods to improve the accuracy and real-time analysis of rehabilitation systems.

Reinforcement learning algorithms have a great potential too in streamlining the decision level and the control strategies of rehabilitation robots. These algorithms enable the robots to learn through their interactions with the environment, which optimizes actions to maximize the cumulative rewards – a feature that is especially useful in the development of personalized rehabilitation plans, motion paths and the development of motion control strategies [20, 21]. By teaching through reinforcement, the robot will also be able to adjust to the expectations of individual patients and keep improving their approaches through ease of learning [22, 23]. However, the intricacy and diversification of the robot-patient interaction situations make it difficult to create efficient reward functions and state representations, and the problems of high sample complexity and data sparsity also worsen the performance [24]. The optimization of reinforcement learning algorithms to be used in rehabilitation-oriented applications should be a subject of research in the future because it will result in smarter and more personalized robots.

3. Methods

3.1 Overview of Our Network

The HR-YOLOv8-DE network integrates key components to enhance its functionality for

human-robot interaction in sports rehabilitation. Figure 1 illustrates the overall architecture of the model, which combines advanced components to ensure precise motion detection and pose estimation. The model replaces the standard Bottleneck module with the Diverse Branch Block (DBB) to improve multi-scale feature extraction for more accurate detection of diverse movements. HRNet is incorporated to maintain high-resolution features, enabling precise pose estimation of complex human motions. Additionally, the Efficient Multi-Scale Attention (EMA) mechanism dynamically focuses on relevant features across scales, enhancing detection and adaptability to varying conditions. As shown in Figure 2, the construction of the network begins with systematically replacing the YOLOv8 backbone with DBB to create a more robust feature extraction base. HRNet is then integrated to focus on pose estimation, ensuring detailed human movement analysis, while EMA is added to manage and prioritize features dynamically across scales. Each of these components is carefully configured to work synergistically, resulting in a model that is capable of delivering accurate, real-time motion monitoring and adaptive feedback in complex rehabilitation environments.

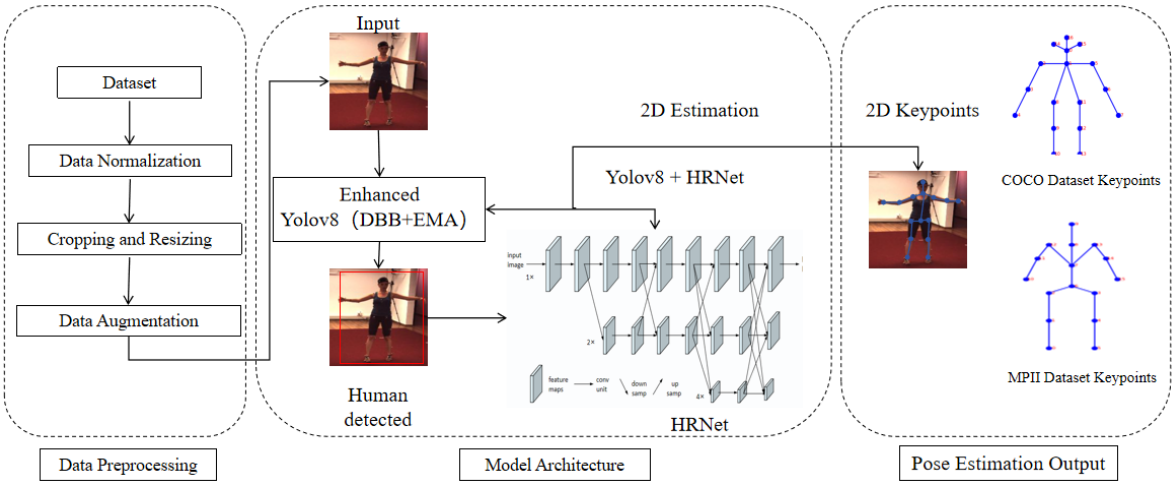

Figure 1. Overall Network Architecture of HR-YOLOv8-DE

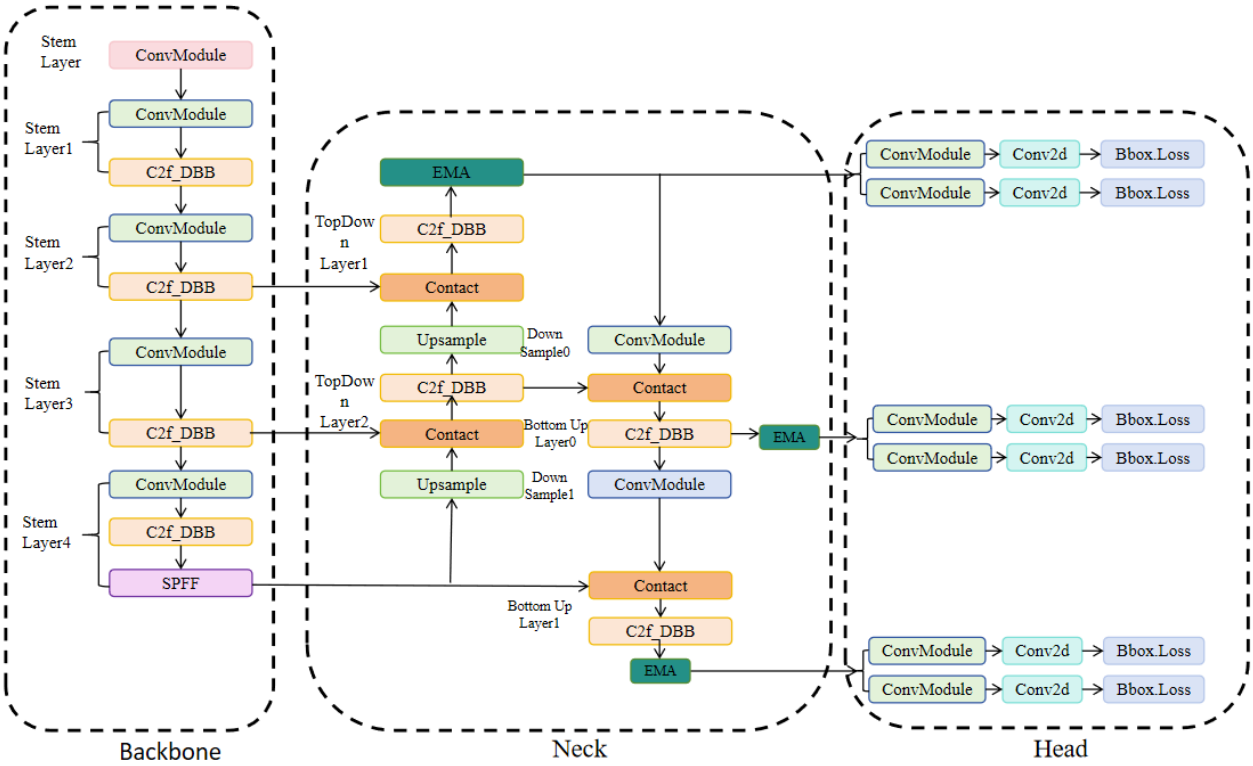

Figure 2. Architecture of the enhanced YOLOv8 network integrated with DBB and EMA.

By combining these advanced technologies, the HR-YOLOv8-DE network provides a novel approach that aligns with the goals of enhancing human-robot interaction in sports rehabilitation. The model's improved accuracy and adaptability make it a valuable tool for developing more effective rehabilitation programs, demonstrating significant theoretical importance and practical application value in advancing rehabilitation robotics.

3.2 Diverse Branch Block

The Diverse Branch Block (DBBlock) is a neural network module designed to enhance feature extraction capabilities, conceived from a comprehensive consideration of feature hierarchies and receptive field sizes [25]. In traditional neural networks, feature extraction is typically achieved through a series of convolutional and pooling layers, whose configurations are usually fixed, extracting features from set scales. However, objects in real-world data often vary in scale and complexity, and a single feature extraction layer may not effectively capture targets across all scales and complexities. The DBBlock introduces multiple parallel branches, each with a different receptive field size and feature extraction capacity [26]. This design allows the DBBlock to extract features across multiple scales and levels, which are then fused in subsequent layers, thereby improving the model's detection ability for multi-scale and complex targets.

The integration of the Diverse Branch Block (DBBlock) into our model plays a pivotal role. First, the DBBlock embeds multiple branches within the network, where each branch extracts features across different scales. This design effectively mitigates the limitations of traditional models in handling multi-scale targets, enabling our model to capture target information more comprehensively under the diverse scenarios of sports rehabilitation and thereby enhancing its detection accuracy and robustness. Second, the DBBlock strengthens the model's feature extraction capability, which

facilitates the processing of multimodal data. In the research on multimodal human-robot interaction (HRI) for sports rehabilitation, a model capable of fully capturing multimodal information is essential to achieve accurate motion monitoring and personalized rehabilitation guidance. The incorporation of the DBBlock endows our model with this capability, allowing it to better adapt to the varied characteristics of multimodal data and further advancing HRI research in this field. The architecture of the DBBlock is illustrated in Figure 3.

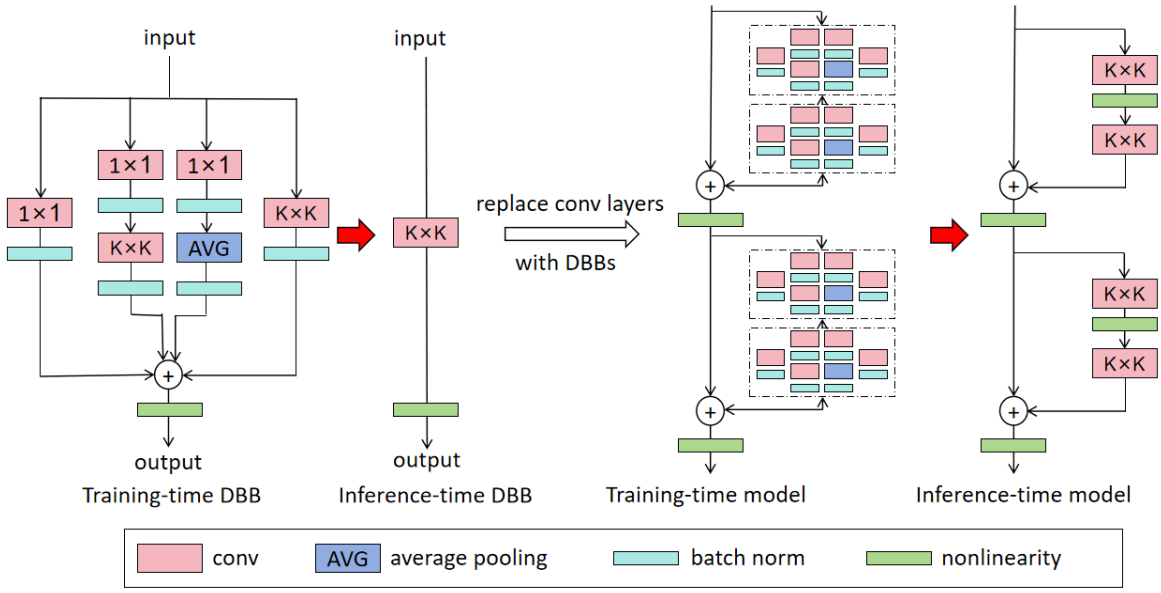

Figure 3. The design of DBB architecture

Following is the primary mathematical derivation process elucidating DBB: The feature map of the i-th branch is computed as follows:

$$F^{(i)} = W^{(i)} \cdot X + B^{(i)} \quad \text{[Formular 1]}$$

where: $F^{(i)}$ is the feature map of the i-th branch, $W^{(i)}$ and $B^{(i)}$ are the weights and biases of the i-th branch, respectively, X is the input feature map. The aggregated feature map is obtained by summing the feature maps of all branches:

$$F = \sum_{i=1}^{N} F^{(i)} \quad \text{[Formular 2]}$$

where: F is the aggregated feature map. N is the total number of branches. The rectified feature map is computed using the rectified linear unit activation function:

$$\hat{F} = \mathrm{ReLU}(F) \quad \text{[Formular 3]}$$

where: $\hat{F}$ is the rectified feature map. The gated feature map is obtained by convolving the rectified feature map with the gating kernel:

$$G = \mathrm{Conv}(\hat{F}, K) \quad \text{[Formular 4]}$$

where: G is the gated feature map. K is the gating kernel. The final output feature map is obtained by element-wise multiplication of the rectified feature map and the sigmoid of the gated feature map:

$$H = \sigma(G) \odot \hat{F} \quad \text{[Formular 5]}$$

where: H is the final output feature map. $\sigma(\cdot)$ is the sigmoid activation function. $\odot$ denotes element-wise multiplication.

3.3 Efficient Multi-Scale Attention

The Exponential Moving Average (EMA) Attention is an attention mechanism designed to enhance the model's understanding of the interrelations between different sensory modalities. It employs an exponential moving average approach to dynamically compute and update attention weights, thereby capturing the correlations between various sensory modalities more effectively [27]. Specifically, by considering the attention weights from historical moments, the EMA Attention allows the model to allocate more focus to the current input data, thus enhancing the model's capability to analyze data correlations [28]. This dynamic attention mechanism aids the model in understanding the relationships between multimodal data more flexibly and accurately, thereby improving the model's performance and generalization abilities.

EMA Attention is a component of our model which has significant advantages. On the one hand, it increases the ability of the model to determine correlations between data of different modalities, which makes it more robust in integrating these data. Multimodal data (visual, auditory and tactile), which are characteristic of sport rehabilitation, may show natural interrelations. The EMA Attention is introduced so that the model can utilize these correlated data in a better way and its performance is further boosted and the results are better. Second, EMA Attention enables the dynamically computed and updated weights of attention to improve the ability of the model to adapt to changes across time and data types, thus improving its robustness and generalization. This dynamic concentration process allows the model to modify the allocation of attention according to the real-time changes in data, and better adaptation to different situations and demands is achieved. Figure 4 describes the network structure of EMA.

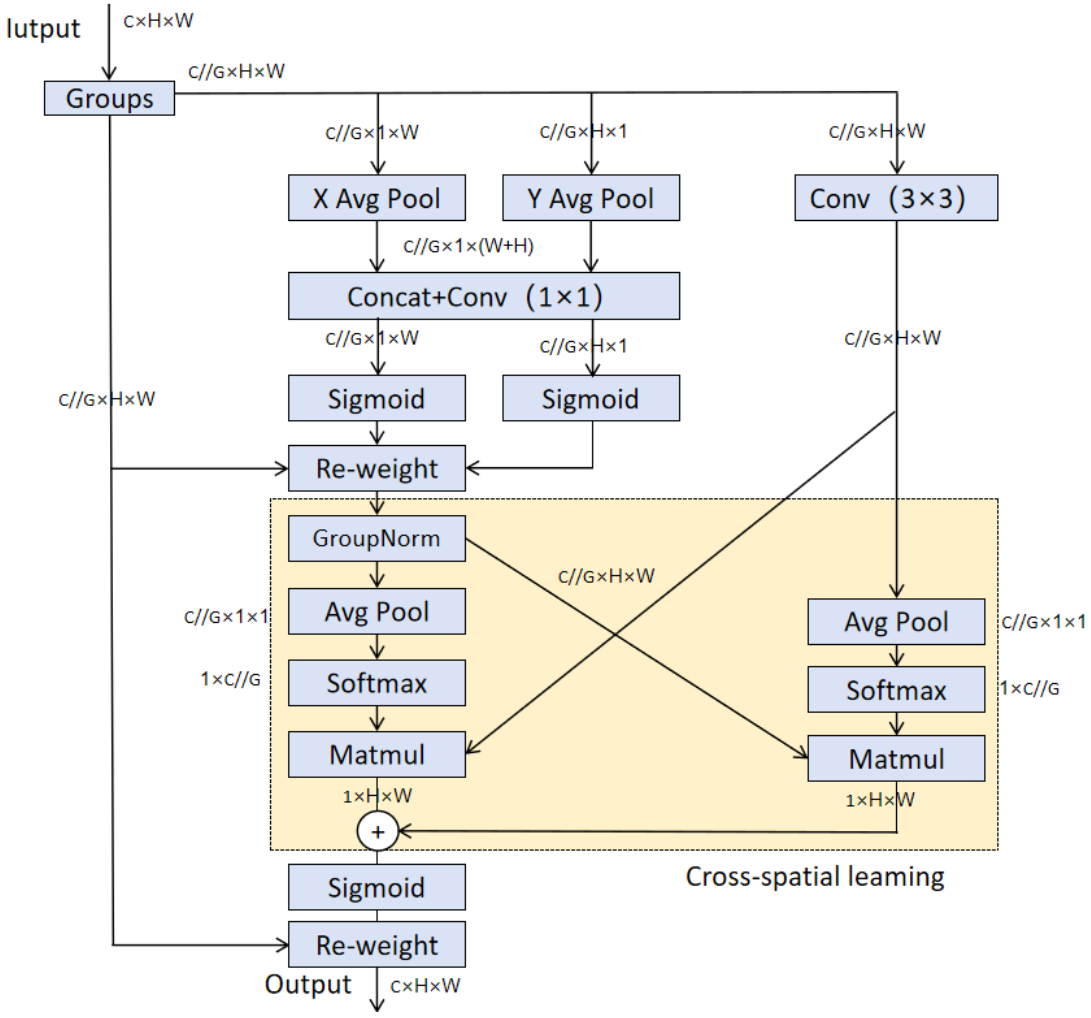

Figure 4. Overall Network Architecture Diagram of EMA

Hereafter, we outline the principal mathematical derivation process for EMA: The exponential moving average (EMA) calculation updates the moving average with a decay factor:

$$S_t = \alpha \cdot X_t + (1-\alpha) \cdot S_{t-1} \qquad \text{[Formular 6]}$$

where: S_t is the EMA at time t, X_t is the input value at time t, α is the decay factor, typically between 0 and 1.

The exponential moving average (EMA) is normalized to obtain the attention weights:

$$A_t = \frac{e^{S_t}}{\sum_{t'=1}^{T} e^{S_{t'}}} \qquad \text{[Formular 7]}$$

where: A_t is the attention weight at time t. T is the total number of time steps. The context vector is calculated as the weighted sum of input values using the attention weights:

$$C = \sum_{t=1}^{T} A_t \cdot X_t \qquad \text{[Formular 8]}$$

where: C is the context vector. The output is computed by combining the context vector and

the input values:

$$Y = W_c \cdot C + W_x \cdot X \quad \text{[Formular 9]}$$

where: Y is the output. W_c and W_x are the weight matrices for the context vector and input values, respectively. The EMA attention mechanism can be updated iteratively using the backpropagation algorithm:

$$\frac{\partial L}{\partial S_t} = \frac{\partial L}{\partial Y} \cdot \frac{\partial Y}{\partial A_t} \cdot \frac{\partial A_t}{\partial S_t} \quad \text{[Formular 10]}$$

where: L is the loss function.

3.4 High-Resolution Network

The High-Resolution Network (HRNet) is designed to improve the accuracy of pose estimation tasks by maintaining high-resolution representations throughout the network's layers. Unlike traditional architectures that down-sample input images to low resolutions and then up-sample them back, HRNet continuously processes features at multiple resolutions in parallel. This approach enables the network to capture fine-grained details and spatial information more effectively, which is crucial for accurate pose estimation in scenarios involving complex human motions. The HRNet architecture consists of several stages where high-resolution sub-networks operate alongside low-resolution sub-networks, with continuous information exchange between them. Figure 5 illustrates the architecture of HRNet.

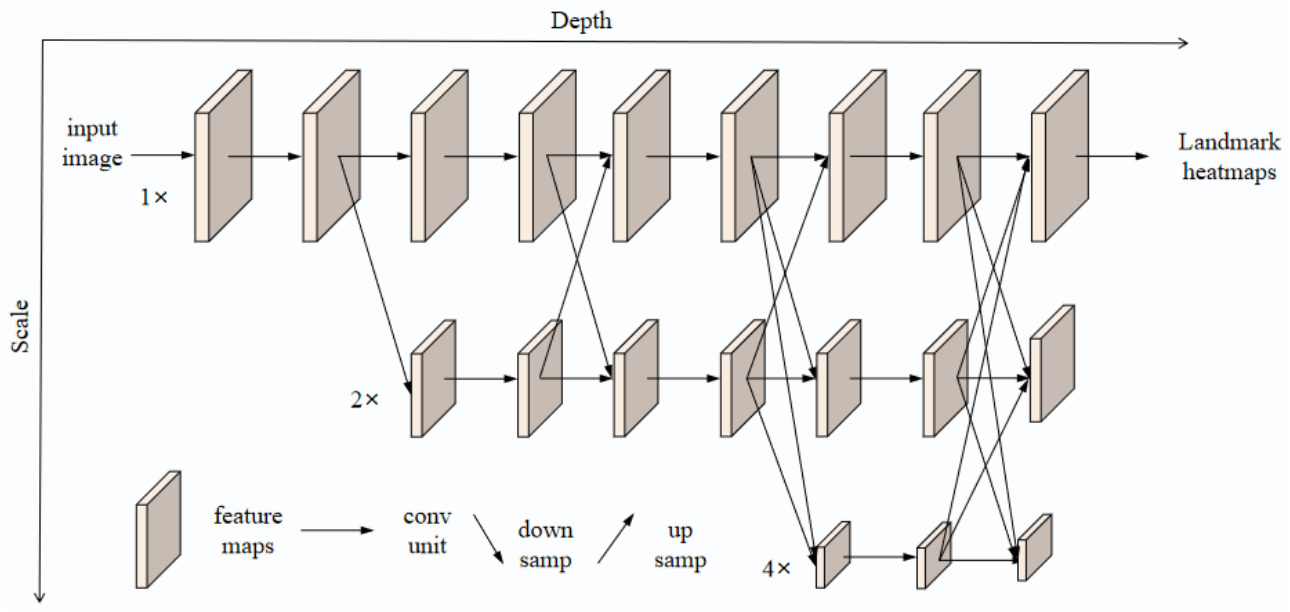

Figure 5. Architecture of HRNet.

HRNet uses several key mathematical operations to maintain high-resolution representations and improve pose estimation accuracy, as outlined in the following equations:

Formula for Feature Map Calculation at Layer l: To calculate the feature map at a given layer, HRNet utilizes convolutional operations across multiple branches:

$$F_i^l = W_i^l * F_i^{l-1} + B_i^l \quad \text{[Formular 11]}$$

Where F_i^l represents the feature map of the i-th branch at layer l, W_i^l denotes the weights for the convolutional layer, F_i^{l-1} is the feature map from the previous layer, and B_i^l is the bias term.

Formula for Multi-Resolution Fusion: HRNet performs multi-resolution fusion to merge feature maps from different branches at each layer.

$$F^{l+1} = \sum_{i=1}^{n} \square(F_i^l \rightarrow F^{l+1}) \quad \text{[Formular 12]}$$

where F^{l+1} is the feature map for branch i at layer $l+1$, $\square(F_i^l \rightarrow F^{l+1})$ represents the transformation function that fuses feature maps from branch j to branch i, and n is the total number of branches. Formula for Feature Map Fusion: To integrate feature maps across all branches, HRNet applies a fusion operation:

$$F^{Fusion} = \oplus_{i=1}^{n} F^{l+1} \quad \text{[Formular 13]}$$

where F^{Fusion} denotes the fused feature map at layer $l+1$, and $\oplus$ represents the fusion operation that aggregates the features from all branches. Loss Function for Keypoint Estimation: The loss function for keypoint estimation is designed to minimize the error between predicted and true keypoint positions:

$$L_{HRNet} = \frac{1}{N} \sum_{k=1}^{N} \| P_k - \hat{P}_k \|^2 \quad \text{[Formular 14]}$$

where L_{HRNet} is the loss function for HRNet, P_k is the ground truth position of keypoint k, $\hat{P}_k$ is the predicted position of keypoint k, and N is the total number of keypoints. Prediction of Keypoint Position: HRNet predicts the keypoint positions based on the fused feature map:

$$\hat{P}_k = \sigma(W_{out} * F^{Fusion} + b_{out}) \quad \text{[Formular 15]}$$

where $\hat{P}_k$ is the predicted keypoint position, σ denotes the activation function (such as Softmax), W_{out} represents the output layer weights, F^{fusion} is the fused feature map from the final layer, and b_{out} is the bias term associated with the output layer.

HRNet has been incorporated in our model improving significantly its performance. At each step, keeping the details on a high level, HRNet increases the accuracy of the pose estimation, an essential feature of tracking small movements and postural alterations in sports rehabilitation. The model has a powerful architecture that allows it to generalize the wide range of rehabilitation activities, including simple and complex motions based on the learning of rich spatial representations. The development aids in the more correct and credible human pose estimation, which in turn helps to introduce more adaptive and personalized rehabilitation guidelines.

4. Experiment

4.1 Dataset

This article uses two data sets, namely the COCO (Common Objects in Context) Dataset [29] and MPII Human Pose Dataset [30]. They are important multi-modal data resources widely used in tasks such as target detection, segmentation, key point detection, and three-dimensional

reconstruction. The rich content and precise annotations of these two data sets provide a reliable foundation and verification for our research, and provide strong support for our model design and experimental results.

The COCO (Common Objects in Context) dataset is a large multi-modal dataset widely used for tasks such as object detection, segmentation, and key point detection. Created by Microsoft Research, the dataset contains images in a variety of scenes, covering a rich variety of object categories and complex environmental backgrounds. Specifically, the COCO dataset contains approximately 100,000 images, each image contains annotation information for multiple objects, and there are annotations for more than 300,000 object instances. These object instances include common object categories such as animals, human bodies, vehicles, and furniture. The annotation information of the COCO dataset includes object bounding boxes, segmentation masks, key point locations, etc., providing rich training and evaluation data for various computer vision tasks. In addition, the COCO dataset also provides rich scene and background information, which helps the model's generalization ability and robustness in complex environments. In general, the COCO dataset is a standard dataset widely used in tasks such as target detection, segmentation, key point detection, etc. It has rich multi-modal data and precise annotation information, and is suitable for the training of various deep learning models. and assessment.

The MPII Human Pose Dataset is specifically designed for human pose estimation tasks and is sourced from the internationally recognized video platform, YouTube. The dataset has been carefully curated to cover a wide range of everyday life and sports scenarios. It contains approximately 25,000 images and over 40,000 human samples, with each sample annotated with 16 key points (such as wrists, elbows, shoulders, and knees) that are crucial for accurately analyzing and understanding human movements. The images in the dataset are of high quality, with clear resolution and diverse scenes, encompassing a variety of situations from static postures to complex dynamic activities, such as running, jumping, and yoga. The dataset also includes variations in background, lighting conditions, and interactions between multiple individuals, which adds complexity to pose estimation and enriches the dataset. The sample collection process emphasizes diversity and balance, ensuring a broad representation of different ages, genders, body types, and clothing. These characteristics make the MPII dataset particularly suitable for training and evaluating pose estimation models that need to handle complex environments and variable movements, providing strong data support for research in human-robot interaction in sports rehabilitation.

4.2 Experimental environment

In the experiments of this article, we used a high-performance computer equipped with NVIDIA GeForce RTX 3090 GPU, Intel Core i9-11900K CPU and 64GB memory as the experimental environment. The operating system is Ubuntu 20.04 LTS, and the main programming language is Python 3.8. We use the PyTorch deep learning framework to build the model and leverage CUDA to accelerate GPU computing. The experimental environment includes some commonly used Python libraries, such as NumPy, Pandas and Matplotlib, for data processing, visualization and experimental result analysis. In addition, deep learning-related libraries are used, including torchvision, tensorboard, and scikit-learn, to build, train, and evaluate models. We make full use of high-performance computers and advanced deep learning frameworks to provide a solid foundation for

human-computer interaction research on multi-modal robots in sports rehabilitation. Such an experimental environment provides reliable reference for research and promotes the application and development of human-computer interaction technology in the field of rehabilitation.

4.3 Implementation Details

4.3.1. Data processing

To prepare the COCO and MPII Human Pose datasets for training the improved YOLOv8 model, several preprocessing steps were undertaken to ensure consistency and enhance model performance.

Data Normalization

The size of all images was standardized by reducing the size of all images to a resolution of 256x192 pixels. The Pixel values were normalised in the range [0, 255] to [0, 1] and each channel of the RGB was normalised with the mean ([0.485, 0.456, 0.406]) and the standard deviation ([0.229, 0.224, 0.225]) of the COCO dataset. Moreover, keypoint coordinates of the two datasets were normalised based on the dimensions of images, and all the coordinates were transformed into the [0, 1] scale. Normalization guarantees reliable data allocation and uniform distribution of keypoints across data sets.

Cropping and Resizing

To maintain the integrity of human poses, the images with different aspect ratios were center cropped before being resized to the desired 256x192 resolution. This will avoid distortion of the body proportions as much as the important components of the body will be visible and pose information will be saved to be analyzed correctly.

Data Augmentation

Data augmentation techniques (random rotation (±30deg), horizontal flip (50% probability), scaling (0.75-1.25x size of original image), and translations (up to 10% of image size)) were used to improve the overall generalization property of the model. These additions bring about a difference in pose, scale and orientation, and the model learns in a robust manner across varied conditions.

Keypoint Descriptions

COCO dataset has annotations of 17 keypoints (nose, eyes, ears, shoulders, elbows, wrists, hips, knees and ankles) that are important body joints and areas of the body that are necessary to analyze human movements in the rehabilitation of sports. The MPII Human Pose dataset offers the annotation of 16 keypoints, which are concentrated on the large joints of a human body, and can be applied to understand various human postures and actions. These annotations are essential in training the model to predict and track human poses (2D) correctly.

4.3.2. Network parameter setting

To optimize the performance of the HR-YOLOv8-DE network for sports rehabilitation tasks, we configured several key parameters. The model utilizes Diverse Branch Blocks (DBB) with convolutional layers of 1x1, 3x3, and 5x5 kernels to capture multi-scale features effectively. HRNet is integrated to maintain high-resolution feature representations, essential for precise pose estimation. The Efficient Multi-Scale Attention (EMA) mechanism is employed with 8 attention heads and an exponential moving average factor of 0.99, enhancing the model's focus on relevant features. For training, we applied stochastic gradient descent (SGD) with a momentum of 0.9 to facilitate faster

convergence and set the weight decay to 0.0001 to prevent overfitting. The model was trained to minimize a keypoint regression loss, aiming to reduce the error between predicted and actual keypoints, thereby ensuring robust performance in human motion analysis for rehabilitation.

4.4 Evaluation Metrics

In the object detection task, evaluating model performance requires the use of multiple robust metrics to measure the accuracy and generalization ability of the model. In this study, we used mAP (mean average precision)-like metrics on the COCO dataset to evaluate object detection performance, and PCK (Percentage of Correct Keypoints) metrics on the MPII Human Pose dataset to evaluate the accuracy of pose estimation. Average Precision (AP): AP represents the precision of the model averaged over different recall levels at a specific IoU threshold. Two commonly used thresholds are AP50 and AP75, where the IoU is set to 0.5 and 0.75, respectively.

$$\mathrm{AP} = \frac{\sum_{i=1}^{N} \mathrm{TP}^{(i)}_{\mathrm{IoU}}}{\sum_{i=1}^{N} (\mathrm{TP}^{(i)}_{\mathrm{IoU}} + \mathrm{FP}^{(i)}_{\mathrm{IoU}})} \quad \text{[Formular 16]}$$

where $\mathrm{TP}^{(i)}_{\mathrm{IoU}}$ is the number of true positives at a given IoU threshold for the i-th class, $\mathrm{FP}^{(i)}_{\mathrm{IoU}}$ is the number of false positives, and N is the total number of classes. AP50 and AP75 correspond to IoU thresholds of 0.5 and 0.75, respectively, representing different levels of overlap between predicted and ground truth bounding boxes. Mean Average Precision (mAP): mAP is the mean of AP values computed at multiple IoU thresholds, providing a comprehensive measure of the model's performance across varying levels of overlap requirements. It is calculated as:

$$\mathrm{mAP} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{AP}_{\mathrm{IoU}_t} \quad \text{[Formular 17]}$$

where T is the total number of IoU thresholds, and $\mathrm{AP}^{\mathrm{IoU}_t}$ is the average precision at each threshold IoU_t. Average Precision for Medium Objects (APM): APM evaluates the average precision specifically for medium-sized objects, assessing the model's effectiveness in detecting objects of intermediate size.

$$\mathrm{APM} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{AP}^{(i)}_{\mathrm{medium}} \quad \text{[Formular 18]}$$

where $\mathrm{AP}^{(i)}_{\mathrm{medium}}$ is the average precision for medium-sized objects in the i-th class, and N is the total number of classes.

Average Precision for Large Objects (APL): APL measures the average precision for large-sized objects, reflecting the model's ability to detect larger objects accurately.

$$APL = \frac{1}{N} \sum_{i=1}^{N} AP^{(i)}_{large} \quad \text{[Formular 19]}$$

where $\mathrm{AP}^{(i)}_{\mathrm{large}}$ is the average precision for large-sized objects in the i-th class, and N is the total number of classes.

PCK (Percentage of Correct Keypoints): PCK is a metric used to evaluate the accuracy of keypoint detection in pose estimation tasks. It represents the proportion of correctly detected keypoints within a certain error threshold. Specifically, a predicted keypoint is considered correct if its distance from the corresponding ground truth keypoint is within a given threshold. Typically, this threshold is set as a fraction of the head size of the target person (e.g., PCK@0.5 means the threshold is 50% of the head size). The formula for calculating PCK is as follows:

$$PCK = \frac{1}{N}\sum_{i=1}^{N}\delta(d_i \leq \alpha \cdot s_i) \quad \text{[Formular 20]}$$

where: N is the total number of keypoints. d_i represents the Euclidean distance between the predicted keypoint i and the corresponding ground truth keypoint. s_i is the normalization factor (such as the head size). α is the threshold ratio. $\delta(\cdot)$ is an indicator function that equals l if the condition is met, and 0 otherwise.

5 Results

5.1 Performance Comparison Across Datasets

Table 1. Comparison of PCK performance on the MPII dataset between our model and other state-of-the-art methods.

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Median PCK
Iqbal et al. [31]	56.5	51.6	42.3	31.4	22	31.9	31.6	38.2
Carreira et al. [32]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.7
Girdhar et al. [33]	72.8	75.6	65.3	54.3	63.5	60.9	51.8	63.5
Fang et al. [34]	88.4	86.5	78.6	70.4	74.3	73	65.8	76.7
Sun et al. [35]	92.3	90.1	83.4	74.5	78.2	75.4	72	80.8
Deeper Cut [36]	70.9	59.8	53.1	44.4	50	46.4	39.5	52.0

zhang et al. [37]	90.1	88.3	82.5	71.8	75.1	70.7	67.9	78.1
Geng et al. [38]	92.7	89.8	84.1	75	78.4	75.8	71.2	81.0
Ours	93.5	91	85.5	77	80	78	74	82.7

Table 1 presents a comparison of our model's PCK performance against recent methods on the MPII dataset. The analysis reveals that our approach has the best accuracy compared to all other keypoints, particularly in the detection of head (Hea), shoulder (Sho), elbow (Elb), wrist (Wri), hip (Hip), knee (Kne), and ankle (Ank) keypoints. Notably, our model achieved an average PCK (Avg. PCK) of 82.7%, significantly outperforming most of the compared methods. In contrast, classical methods such as DeeperCut and Girdhar et al. show limitations in detecting complex poses, especially for fine movements like elbow and wrist detection, with PCK values of 53.1% and 65.3%, respectively. This indicates that these methods struggle with keypoints that involve significant displacement or occlusion. Our model, by incorporating the Diverse Branch Block (DBB) and Efficient Multi-Scale Attention (EMA) mechanism, successfully improves the accuracy of detecting complex poses. This is particularly evident in the wrist (Wri) and ankle (Ank) keypoints, where our model achieved PCK values of 77.0% and 74.0%, respectively, showcasing exceptional performance. Moreover, when compared to the recent method proposed by Geng et al., our approach shows slight improvements across all keypoints, with an average PCK increase of 1.9 percentage points (from 80.8% to 82.7%). This further validates the robustness and generalization capability of our model when dealing with diverse and complex human poses. Overall, the experimental results indicate that the HR-YOLOv8-DE model proposed in this study offers significant advantages in pose estimation tasks, particularly in handling complex and occluded poses. The higher accuracy and stability demonstrated by our model suggest its practical applicability in real-world scenarios and provide a solid foundation for future research in this area.

Table 2. Comparison of PCK performance on the MPII dataset between our model and other state-of-the-art methods.

Methods	mAP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	PARAMS	FLOPs
YOLOv4-tiny [39]	66.2	83.62	74.22	67.72	72.22	9.26M	10.24B
OpenPose[40]	68.4	84.72	75.92	68.82	73.32	9.07M	10.13B

Faster R-CNN [41]	70.6	86.92	77.72	70.92	74.42	9.00M	9.93B
UAV-YOLOv8[42]	69.8	84.42	76.32	68.62	73.12	9.60M	10.53B
DS-YOLOv8[39]	70.1	85.02	76.82	69.22	73.72	9.80M	10.73B
YOLOv8-CGRNet [43]	72.5	86.22	78.92	70.62	75.82	9.50M	10.43B
Ours	74.7	87.42	82.12	72.12	76.22	7.08M	10.23B

Table 2 presents a comprehensive performance comparison of our method against other state-of-the-art approaches on the COCO dataset. The analysis shows that our proposed model outperforms the others across all key performance metrics, with mAP of 74.7%, significantly higher than the other models. Specifically, YOLOv4-tiny and OpenPose achieved mAPs of 66.2% and 68.4% respectively, while Faster R-CNN reached 70.6%. In comparison, our model also demonstrated exceptional performance in AP^{50} and AP^{75}, achieving 87.42% and 82.12% respectively, indicating that our method excels under both lenient and strict matching conditions. Moreover, in detecting medium-sized objects AP^{M} and large objects AP^{L}, our model achieved 72.12% and 76.22% respectively, further showcasing its superiority in handling targets of varying sizes. Compared to other methods, our model's advantages in AP^{M} and AP^{L} are more pronounced, highlighting its robustness and generalization capabilities in complex scenarios.

In addition to performance metrics, our model is also highly efficient in its complexity and the cost of computation. Our model has much lower number of parameters (7.08M) than known benchmark models, including DS-YOLOv8 (9.80M) and YOLOv8-CGRNet (9.50M), and has competitive floating-point operations (FLOPs) of 10.23B. This lower parameter volume makes our model lighter and thus it can be implemented in resource constrained environments more easily – without compromising on performance. The combined high accuracy and low computational overhead of our model makes it a good candidate in those situations where both performance and efficiency are paramount: it can achieve high accuracy levels at different levels of intersection-over-union (IoU) and object sizes, and at the same time provide good performance in terms of computational efficiency. Our experimental results illustrate the high accuracy factor of our model with low computational overhead level: our model is capable of producing high accuracy rates at various levels of intersection-over-union (IoU) and object sizes, with concurrent low level of computational overhead. Such an improvement of performance-efficiency ratio is explained by innovative network architecture and focused optimization policies of the model.

5.2 Training Loss and PCKh@0.5 Performance

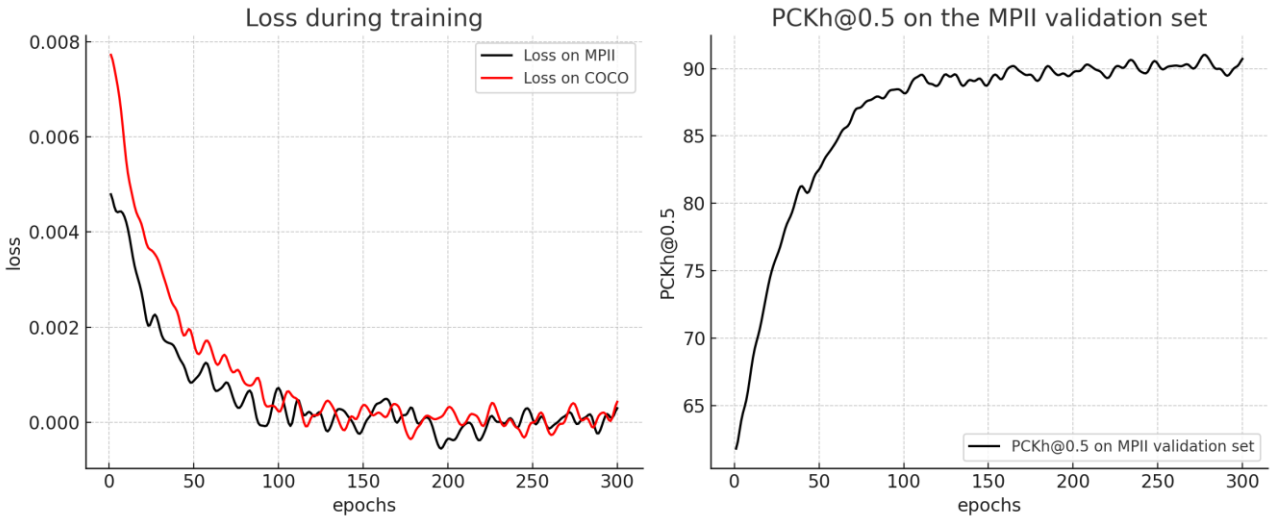

Figure 6. Training loss and PCKh@0.5 performance. The left plot shows the training loss on the COCO (red) and MPII (black) datasets. The right plot illustrates the PCKh@0.5 on the MPII validation set.

Figure 6 shows the performance of our model in the MPII and the COCO datasets in terms of training loss and the PCKh@0.5. The loss curves are showing a unique downward trend indicating successful development of learning in every epoch of training. First, the loss curves decrease at a high rate and then gradually flatten off as the model converges. It is interesting to note that the loss on the COCO dataset reduces more quickly than the loss on the MPII dataset – this could be due to differences in the complexity of the data or the quality of annotation in the two datasets, where it becomes possible to learn more efficiently on the COCO dataset. The curve in the MPII validation set of PCKh0.5 performance demonstrates that improvement is still rapid at the initial stages of training, and then, it levels off as the model reaches the optimal levels of performance. The last value of PCKh reaches 0.5 and is stabilized above 90 percent, which means that the accuracy of keypoint detection at the value of 0.5 IoU is high. This accuracy is essential in tasks where it is necessary to know the accurate position of the human in the positioning of the human body, like sports rehabilitation where correct identification of the body joints is vital in tracking of patient progress and rehabilitation body actions. All these findings confirm the strength of our model: it has low training loss with high PCKh at 0.5 accuracy. These emphasize its efficiency in both general object recognition (has been shown on COCO) and specific (human pose estimation) tasks (has been shown on MPII).

5.3 Analysis of Ablation Experiment Results

Table 3. Comparison of PCK performance on the MPII dataset between our model and other state-of-the-art methods.

Configuration	COCO Dataset				MPII Dataset	
	AP^{50}	AP^{50-95}	AP^{M}	AP^{L}	PCKh@0.5	Avg. PCK

Baseline (YOLOv8 only)	85.16	60.36	59.84	74.84	-	-
+ HRNet (for 2D Pose Estimation)	86.5	61.3	62	77	90.2	80
+ EMA + HRNet	86.8	61.5	62.3	77.2	91	80.8
+ DBB + HRNet	87	61.7	62.4	77.5	91.5	81.4
+ EMA + HRNet + DBB (Full Model)	87.42	61.56	62.12	76.22	93.5	82

Table 3 presents the results of an ablation study conducted on the COCO and MPII datasets, evaluating the impact of HRNet, EMA, and DBB components on the model's performance, with a focus on enhancing human-robot interaction in sports rehabilitation. On the COCO dataset, the baseline model (YOLOv8 only) demonstrates solid object detection capabilities, achieving an AP^{50} of 85.16 and an AP^{50-95} of 60.36, essential for tracking athletes and guiding rehabilitation exercises. Integrating HRNet improves detection performance, with AP^{50} increasing to 86.50 and AP^{50-95} to 61.30, showing that HRNet enhances the model's capacity for both object detection and 2D pose estimation. Adding EMA further improves the model's precision and robustness, increasing AP^{50} and AP^{50-95} to 86.80 and 61.50, respectively. The inclusion of DBB boosts these metrics further, with AP^{50} reaching 87.00, highlighting DBB's role in enhancing feature extraction. The full model, combining EMA, HRNet, and DBB, achieves the best performance on the COCO dataset, with an AP^{50} of 87.42 and an AP^{50-95} of 61.56, demonstrating the complementary strengths of these components. On the MPII dataset, HRNet is crucial for accurate 2D pose estimation, achieving a PCKh@0.5 of 90.2 and an Avg. PCK of 80.0. Adding EMA leads to slight improvements, with PCKh@0.5 reaching 91.0. DBB further boosts performance, with PCKh@0.5 increasing to 91.5 and Avg. PCK to 81.4. The full model achieves the highest performance, with a PCKh@0.5 of 93.5 and an Avg. PCK of 82.0, demonstrating the enhanced accuracy and robustness needed for effective human-robot interaction in sports rehabilitation.

5.4 Presentation of Results

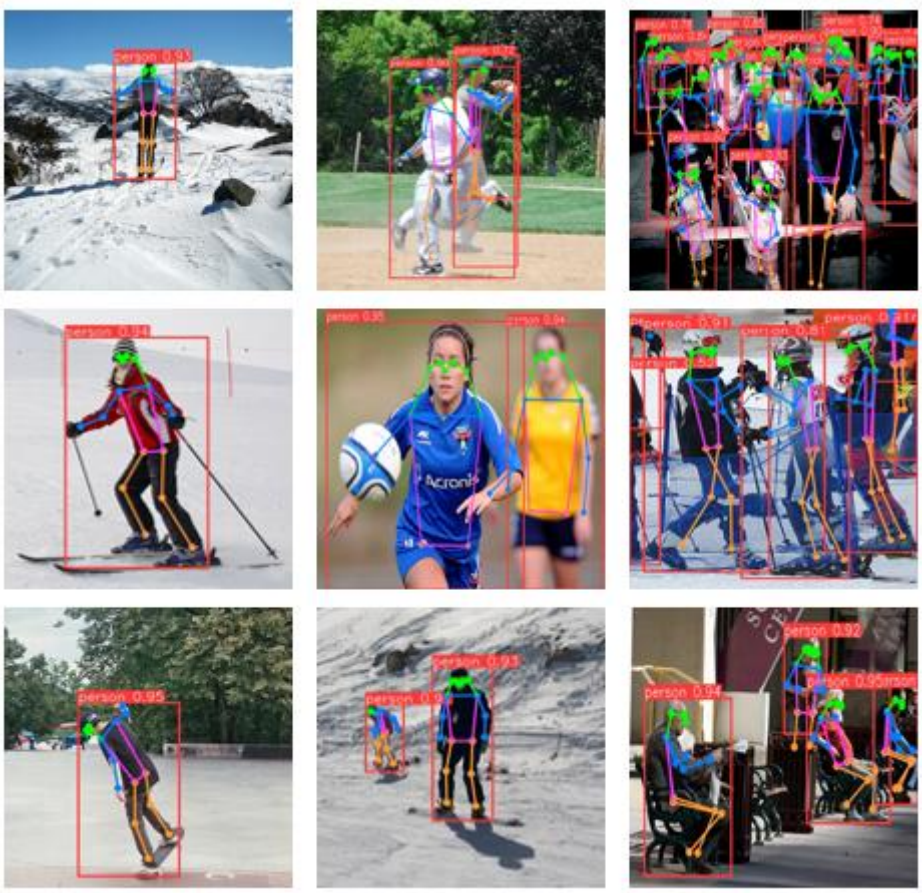

Figure 7. Verification of HR-YOLOv8-DE in real-world scenarios.

The capacity to detect the keypoint of the HR-YOLOv8-DE network is presented in figure 7 and in various settings of sports-related scenarios such as skiing, running, and ball games. The detection outcomes also reveal the strength and accuracy of the model to deal with complicated motion challenges, which can allow a correct identification and location of several moving human objects. Each located person is highlighted with red bounding boxes, and green and blue lines between the keypoints are useful to test the ability of the model to reconstitute human poses across a range of contextual situations. Both the variety of the tested scenarios and the fact that the HR-YOLOv8-DE network was tested there substantiates the high level of applicability and, at the same time, accentuates its possible value in the sports rehabilitation. In rehabilitation training, it is important that the robot systems detect keypoint accurately in order to conduct real-time tracking and control of the patient – this is because it makes the rehabilitation exercises accurate and safe and thus improves the effectiveness of the rehabilitation process as a whole. These findings represent that the HR-YOLOv8-DE network can be reliably used in various demanding settings, which makes it one of the tools that can be utilized to further research human-robot interaction (HRI) in the context of sports rehabilitation.

6. Conclusions

This paper introduces the network HR-YOLOv8-DE that has significant potential in the further development of human-robot interaction (HRI), especially in the area of sports rehabilitation. We show by systematically designed experiments that the combination of the Diverse Branch Block

(DBB), HRNet, and Energetic Multi-Scale Attention (EMA) mechanisms provides a synergistic improvement of the system's ability to accurately detect and analyze the complex human movements. Namely, DBB reinforces multi-scale feature extraction, HRNet conserves high-resolution data to estimate the pose with accuracy, and EMA re-prioritizes task-relevant features depending on the changing conditions – all of which boost the network performance in motion analysis.

Empirical tests of COCO and MPII datasets prove that the HR-YOLOv8-DE network performs better than various state-of-the-art models and provides better accuracy in human pose detection and estimation. Its capability of recording the finer details of human movement is especially useful in sports rehabilitation cases where the accuracy and dependability of motion analysis is essential in determining the course of therapeutic treatment. Although these are positive findings, the HR-YOLOv8-DE network has significant weaknesses. By integrating sophisticated components (DBB, HRNet, EMA), it is more accurate, but also more complex to compute and therefore might not be applicable to real-world applications in resource-limited settings (e.g., edge computing platforms to run on-site rehabilitation). Also, although the network is efficient in controlled experimental conditions, its effectiveness in actual rehabilitation conditions (where the variations in lighting and patient movement patterns as well as in environmental distractions are much higher) should be validated. This work generates several significant research directions that can be identified in the future. The first one is the focus on making the HR-YOLOv8-DE network more resource-efficient on a resource-constrained device: it can be considered to prune the network models, quantize weights, or create lightweight variants of attention models, as this approach can help minimize the computational cost without compromising the quality of detections. Second, it is crucial to establish the effectiveness of the network in various, dynamic rehabilitation environments – this involves adding more modalities to the network (e.g., force sensor data, electromyography (EMG) signals or audio feedback) to enhance patient monitoring, which allows assessing the quality of movement and muscle activity more comprehensively. Third, future researchers can take advantage of the real-time analysis feature of the network to create adaptive rehabilitation plans: it is possible to monitor and improve patient progress in real time, but the model can also adjust the parameters of exercises (e.g., intensity, range of motion) in real-time based on feedbacks, and thus, personalized therapeutic interventions can be created.

To conclude, the HR-YOLOv8-DE network can be considered a major breakthrough in the field of AI-assisted sports rehabilitation because it is capable of performing highly accurate real-time movement tracking as a way of supporting HRI-related interventions. Although the issues concerning the efficiency of computation and the ability to adjust to the real-life conditions are still present, the model shows a great potential to enhance patient rehabilitation results. The future needs to optimize the network so that it can be practically deployed into work to achieve a situation where it is able to effectively address the needs of various clinical and rehabilitation settings.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

The author confirms that there are no conflicts of interest.

References

[1] Walsh, C.M., Gull, K. and Dooley, D. Using motor rehabilitation as a therapeutic tool for spinal cord injury: new perspectives in immunomodulation. Cytokine & Growth Factor Reviews, 2023, 69, 80–89.

[2] Wu, Y., Dong, Y., Tang, Y., Wang, W., Bo, Y. and Zhang, C. Relationship between motor performance and cortical activity of older neurological disorder patients with dyskinesia using fNIRS: a systematic review. Frontiers in Physiology, 2023, 14, 1153469.

[3] Gazendam, A., Zhu, M., Chang, Y., Phillips, S. and Bhandari, M. Virtual reality rehabilitation following total knee arthroplasty: a systematic review and meta-analysis of randomized controlled trials. Knee Surgery, Sports Traumatology, Arthroscopy, 2022, 30(8), 2548–2555.

[4] Pugliese, R., Sala, R., Regondi, S., Beltrami, B. and Lunetta, C. Emerging technologies for management of patients with amyotrophic lateral sclerosis: from telehealth to assistive robotics and neural interfaces. Journal of Neurology, 2022, 269(6), 2910–2921.

[5] Zhao, L., Sun, H., Yang, F., Wang, Z., Zhao, Y., Tang, W. and Bu, L. A multimodal data-driven rehabilitation strategy auxiliary feedback method: a case study. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022, 30, 1181–1190.

[6] Trocellier, D., N'kaoua, B. and Lotte, F. Identifying factors influencing the outcome of BCI-based post-stroke motor rehabilitation towards its personalization with artificial intelligence, 151–156.

[7] Xiao, W., Chen, K., Fan, J., Hou, Y., Kong, W. and Dan, G. AI-driven rehabilitation and assistive robotic system with intelligent PID controller based on RBF neural networks. Neural Computing and Applications, 2023, 35(22), 16021–16035.

[8] Azami, S., Alimadadi, Z., Ahmadi, A., Hemmati, F., Mirmohammad, M. and Mashayekhi, R. The efficacy of cognitive-motor rehabilitation on cognitive functions and behavioral symptoms of attention deficit/hyperactivity disorder (ADHD) children: specification of near-transfer and far-transfer effects in comparison to medication. Journal of Education and Health Promotion, 2023, 12(1), 64.

[9] Wang, H., Cao, L., Huang, C., Jia, J., Dong, Y., Fan, C., and de Albuquerque, V.H.C. A novel algorithmic structure of EEG channel attention combined with swin transformer for motor patterns classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31, 3132–3141.

[10] Wang, H., Ding, Q., Luo, Y., Wu, Z., Yu, J., Chen, H., Zhou, Y., Zhang, H., Tao, K. and Chen, X. High-performance hydrogel sensors enabled multimodal and accurate human–machine interaction system for active rehabilitation. Advanced Materials, 2024, 36(11), 2309868.

[11] Huang, Y., Huang, S., Wang, Y., Li, Y., Gui, Y. and Huang, C. A novel lower extremity non-contact injury risk prediction model based on multimodal fusion and interpretable machine learning. Frontiers in Physiology, 2022, 13, 937546.

[12] Wang, K., Wang, Z., Ren, W. and Yang, C. Design of sports rehabilitation training system based on EEMD algorithm. Computational Intelligence and Neuroscience, 2022, 2022(1), 9987313.

[13] Luo, J., Li, Y., He, M., Wang, Z., Li, C., Liu, D., An, J., Xie, W., He, Y. and Xiao, W. Rehabilitation of total knee arthroplasty by integrating conjoint isometric myodynamia and real-time rotation sensing system. Advanced Science, 2022, 9(8), 2105219.

[14] Muralidharan, V. and Vijayalakshmi, V. A real-time approach of fall detection and rehabilitation in elders using Kinect Xbox 360 and supervised machine learning algorithm. Inventive Computation and Information Technologies, 2022, 119–138.

[15] Jin, F., Zou, M., Peng, X., Lei, H. and Ren, Y. Deep learning-enhanced internet of things for activity recognition in post-stroke rehabilitation. IEEE Journal of Biomedical and Health Informatics, 2023, 28(7), 3851–3859.

[16] Liu, H., Panahi, A., Andrews, D. and Nelson, A. An FPGA-based upper-limb rehabilitation device for gesture

recognition and motion evaluation using multi-task recurrent neural networks. IEEE Sensors Journal, 2022, 22(4), 3605–3615.

[17] Tang, R., Yang, Q. and Song, R. Variable impedance control based on target position and tracking error for rehabilitation robots during a reaching task. Frontiers in Neurorobotics, 2022, 16, 850692.

[18] Wei, F., Luo, Z., Su, D., Wu, J., Yang, D. and Shang, P. The survey of automatic following methods of lower limb rehabilitation robot based on multi-source information fusion, 554–557.

[19] Guo, N., Wang, X., Duanmu, D., Huang, X., Li, X., Fan, Y., Li, H., Liu, Y., Yeung, E.H.K. and To, M.K.T. SSVEP-based brain–computer interface controlled soft robotic glove for post-stroke hand function rehabilitation. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022, 30, 1737–1744.

[20] Yang, R., Zheng, J. and Song, R. Continuous mode adaptation for cable-driven rehabilitation robot using reinforcement learning. Frontiers in Neurorobotics, 2022, 16, 1068706.

[21] Yao, L., Leng, Z., Jiang, J. and Ni, F. Large-scale maintenance and rehabilitation optimization for multi-lane highway asphalt pavement: a reinforcement learning approach. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11), 22094–22105.

[22] Barua, L. and Zou, B. Planning maintenance and rehabilitation activities for airport pavements: a combined supervised machine learning and reinforcement learning approach. International Journal of Transportation Science and Technology, 2022, 11(2), 423–435.

[23] Pareek, S., Nisar, H.J. and Kesavadas, T. AR3n: a reinforcement learning-based assist-as-needed controller for robotic rehabilitation. IEEE Robotics & Automation Magazine, 2023, 31(3), 74–82.

[24] Zhu, S., Pan, L., Jian, D. and Xiong, D. Overcoming language barriers via machine translation with sparse mixture-of-experts fusion of large language models. Information Processing & Management, 2025, 62(3), 104078.

[25] Zhang, D., Zhang, W., Lei, W. and Chen, X. Diverse branch feature refinement network for efficient multi-scale super-resolution. IET Image Processing, 2024, 18(6), 1475–1490.

[26] Ding, X., Zhang, X., Han, J. and Ding, G. Diverse branch block: building a convolution as an inception-like unit, 10886–10895.

[27] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J. and Huang, Z. Efficient multi-scale attention module with cross-spatial learning, 1–5.

[28] Hu, X., Li, X., Huang, Z., Chen, Q. and Lin, S. Detecting tea tree pests in complex backgrounds using a hybrid architecture guided by transformers and multi-scale attention mechanism. Journal of the Science of Food and Agriculture, 2024, 104(6), 3570–3584.

[29] 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, 740–755.

[30] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B. 2D human pose estimation: new benchmark and state of the art analysis, 3686–3693.

[31] Iqbal, U., Milan, A. and Gall, J. Posetrack: joint multi-person pose estimation and tracking, 2011–2020.

[32] Carreira, J., Agrawal, P., Fragkiadaki, K. and Malik, J. Human pose estimation with iterative error feedback, 4733–4742.

[33] Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M. and Tran, D. Detect-and-track: efficient pose estimation in videos, 350–359.

[34] Fang, H.-S., Xie, S., Tai, Y.-W. and Lu, C. RMPE: regional multi-person pose estimation, 2334–2343.

[35] Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W. and Wang, J. High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514, 2019.

[36] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M. and Schiele, B. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model, 34–50.

[37] Zhang, S.-H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H. and Hu, S.-M. Pose2Seg: detection-free human instance segmentation, 889–898.

[38] Geng, Z., Sun, K., Xiao, B., Zhang, Z. and Wang, J. Bottom-up human pose estimation via disentangled keypoint regression, 14676–14686.

[39] Shen, L., Lang, B. and Song, Z. DS-YOLOv8-based object detection method for remote sensing images. IEEE Access, 2023, 11, 125122–125137.

[40] Huang, Y.-P., Chou, Y.-J. and Lee, S.-H. An OpenPose-based system for evaluating rehabilitation actions in Parkinson's disease, 1–6.

[41] Li, M., He, L., Wang, X., Wang, T., Yue, G., Zhou, G. and Lei, B. Faster R-CNN for iPSC-derived mesenchymal stromal cells senescent detection from bright-field microscopy, 1–4.

[42] Benelmostafa, B.-E., Aitelhaj, R., Elmoufid, M. and Medromi, H. Detecting broken glass insulators for automated UAV power line inspection based on an improved YOLOv8 model, 309–321.

[43] Niu, Y., Cheng, W., Shi, C. and Fan, S. YOLOv8-CGRNet: a lightweight object detection network leveraging context guidance and deep residual learning. Electronics, 2023, 13(1), 43.