INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 01 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

OpenPose EfficientDet Feedback Supports Upper Limb Rehabilitation Robots in Clinical Motion Training with Environmental Cues in Clinics

Lydia Bennett, Mathieu Rolland, Serena Ferrari, Federica Ferrara

Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia

Abstract *Neurological disorders, particularly stroke, represent a leading cause of long-term disability worldwide, heavily impacting upper-limb motor function and diminishing the overall quality of life for millions of individuals. Traditional upper-limb rehabilitation relies extensively on manual therapy, which is highly labor-intensive, difficult to quantify, and restricted by the availability of specialized physical therapists. In response to these challenges, robotic rehabilitation has emerged as a promising adjunct to clinical practice. However, many current robotic systems lack real-time, context-aware visual feedback mechanisms that can accurately interpret both patient kinematics and interaction with therapeutic objects. This paper proposes a novel framework integrating the OpenPose human pose estimation algorithm and the EfficientDet object detection architecture to provide comprehensive motion and environmental feedback for upper-limb rehabilitation robots in clinical settings. By combining markerless skeletal tracking with high-efficiency object recognition, the proposed system enables robots to adaptively adjust task difficulty and provide corrective guidance based on the precise spatial relationship between the patient and the target objects. The methodology outlines the integration of these two deep learning pipelines, detailing the synchronization and data fusion processes required for low-latency clinical applications. Experimental validation in a controlled clinical environment demonstrates significant improvements in kinematic tracking accuracy, therapeutic object interaction detection, and overall system responsiveness. Furthermore, qualitative assessments suggest enhanced patient engagement and adherence due to the immediate, visually driven corrective feedback. The findings indicate that the unified OpenPose-EfficientDet framework offers a robust, scalable, and highly effective solution for advancing intelligent robotic rehabilitation therapies.*

Keywords: *Upper-Limb Rehabilitation, OpenPose, EfficientDet, Robotic Feedback*

1 Introduction

1.1 Background and Clinical Context

The global burden of neurological disorders has reached unprecedented levels, with stroke remaining a predominant cause of acquired adult disability across both developed and developing nations [1]. Among stroke survivors, upper-limb hemiparesis is one of the most common and debilitating consequences, severely restricting the ability to perform basic

activities of daily living such as feeding, dressing, and personal hygiene [2]. The restoration of upper-limb function is therefore a primary focus of neurorehabilitation [3]. The foundational principle underlying motor recovery is neuroplasticity, which refers to the brain's inherent capacity to reorganize its neural pathways in response to intensive, repetitive, and task-specific motor training [4]. Historically, this training has been delivered through manual therapy administered by trained physical therapists. While effective, manual therapy is constrained by several critical limitations [5]. It is highly labor-intensive, physically demanding for the therapists, and often lacks the objective, high-resolution quantitative data necessary to accurately track a patient's micro-progressions over time [6]. Furthermore, the growing aging population has created a significant mismatch between the high demand for rehabilitation services and the limited supply of clinical professionals, necessitating the development of scalable, automated therapeutic solutions [7]. Robotic rehabilitation systems have been increasingly integrated into clinical workflows to address these systemic challenges [8]. End-effector and exoskeleton-based robotic devices offer the ability to deliver highly repetitive, consistent, and intensive motor training without succumbing to fatigue [9]. These systems can operate in various modes, including passive, active-assisted, and resistive, tailoring the physical interaction to the specific impairment level of the patient [10]. Despite their mechanical sophistication, a major limitation of early-generation rehabilitation robots is their reliance on purely kinesthetic or force-based feedback mechanisms, often neglecting the rich visual and contextual information present in the clinical environment [11]. Traditional systems typically rely on physical encoders and load cells to determine the position of the robotic arm, but they remain fundamentally unaware of the patient's holistic body posture, compensatory movements, or dynamic interactions with external therapeutic objects [12].

1.2 The Need for Advanced Visual Feedback

To maximize the efficacy of robotic rehabilitation, it is essential to incorporate advanced sensory modalities that provide real-time, context-aware feedback [13]. Visual feedback, facilitated by computer vision technologies, offers a non-intrusive method for capturing high-dimensional data regarding the patient's kinematics and their interaction with the environment [14]. In clinical practice, physical therapists constantly observe the patient to detect compensatory movements, such as trunk leaning or shoulder elevation, which patients frequently employ to overcome distal arm weakness [15]. If left uncorrected, these compensatory strategies can lead to maladaptive neuroplasticity and secondary musculoskeletal complications, ultimately hindering true motor recovery [16]. Early attempts to integrate visual tracking into rehabilitation robotics relied heavily on marker-based optical motion capture systems [17]. While these systems offer sub-millimeter accuracy and are considered the gold standard in biomechanical analysis, they are highly impractical for routine clinical use [18]. Marker-based systems require expensive, specialized camera arrays, extensive calibration procedures, and the time-consuming attachment of physical markers directly onto the patient's skin or clothing [19]. This setup time significantly reduces the actual time available for active therapy and can cause discomfort or anxiety for patients, particularly those with sensory processing sensitivities [20]. Consequently, there has been a paradigm shift towards markerless motion capture technologies driven by recent advancements in deep learning and convolutional neural networks [21].

1.3 Introduction to OpenPose and EfficientDet

The proposed framework leverages two state-of-the-art deep learning architectures: OpenPose and EfficientDet. OpenPose represents a breakthrough in markerless human pose estimation, utilizing a bottom-up approach to simultaneously detect multiple individuals and their corresponding joint keypoints in real-time [22]. By extracting the spatial coordinates of essential anatomical landmarks such as the shoulder, elbow, and wrist, OpenPose enables the continuous calculation of joint angles and kinematic trajectories without the need for physical markers [23]. This capability is crucial for identifying pathological movement patterns and providing corrective feedback to the robotic control system [24].

However, monitoring the patient's kinematics in isolation is insufficient for functional, task-oriented rehabilitation [25]. Functional therapy often involves the manipulation of physical objects, such as blocks, cups, or specialized therapeutic tools, to simulate activities of daily living [26]. The robotic system must not only understand where the patient's arm is but also where the target objects are located and how the patient is interacting with them [27]. To address this requirement, the framework incorporates EfficientDet, a highly scalable and computationally efficient object detection model [28]. EfficientDet utilizes a bidirectional feature pyramid network and a customized scaling compound to achieve exceptional accuracy in identifying and localizing objects within the camera's field of view, even in cluttered clinical environments [29]. By fusing the kinematic data generated by OpenPose with the environmental awareness provided by EfficientDet, the proposed system creates a comprehensive, closed-loop feedback mechanism [30]. The rehabilitation robot can dynamically adjust its level of physical assistance, modify task trajectories, and provide targeted visual or auditory cues based on a holistic understanding of the patient-object interaction [31]. This paper details the architectural design, algorithmic integration, and clinical evaluation of this combined OpenPose-EfficientDet motion feedback system, demonstrating its potential to significantly enhance the efficacy and adaptability of upper-limb rehabilitation robotics.

2. Literature Review and Background

2.1 Evolution of Robotic Rehabilitation

The integration of robotics into neurorehabilitation has evolved significantly over the past three decades. The primary objective of these systems has always been to replicate the intensive, repetitive movements required to drive neuroplasticity, while simultaneously providing objective data on patient performance [32]. The first generation of rehabilitation robots, such as the MIT-Manus, focused on planar movements in a horizontal workspace [33]. These end-effector devices proved highly effective in improving motor control in the chronic stages of stroke, demonstrating that robotic therapy could match or exceed the outcomes of conventional manual therapy when matched for intensity [34]. Subsequent developments introduced highly articulated exoskeletons designed to encompass the entire upper limb, mimicking human anatomical degrees of freedom. Exoskeletons provide precise control over individual joint torques and allow for three-dimensional, spatial reaching tasks [35]. However, the mechanical complexity of exoskeletons poses challenges in alignment, calibration, and patient comfort. Moreover, both end-effector and exoskeleton systems historically relied on internal sensors to measure progress, which limited their ability to perceive the patient's global posture and the surrounding clinical context. This limitation spurred the exploration of external sensing modalities to complement the internal robotic data [36].

2.2 Computer Vision in Clinical Assessment

The application of computer vision in clinical assessment and rehabilitation has grown exponentially alongside the rapid advancements in artificial intelligence. Initially, vision-based systems were confined to simple tasks such as tracking color-coded objects or utilizing infrared depth sensors like the Microsoft Kinect for gross motor evaluation [37]. While the Kinect popularized the concept of accessible, markerless tracking in gaming and home rehabilitation, its proprietary skeletonization algorithms struggled with occlusions, atypical postures common in neurological patients, and complex backgrounds found in clinical settings [38]. The advent of robust convolutional neural networks revolutionized the field of visual tracking. Researchers began employing deep learning models to extract intricate anatomical features from standard two-dimensional RGB video streams [39]. These approaches eliminated the need for specialized depth hardware and significantly improved robustness against environmental variations [40]. In the context of upper-limb rehabilitation, accurate pose estimation is critical for evaluating the smoothness, straightness, and coordination of reaching movements, which are standard clinical metrics for motor recovery.

2.3 Markerless Kinematic Tracking with OpenPose

OpenPose stands out as a pioneering architecture in the domain of markerless human pose estimation. Unlike top-down approaches that first detect bounding boxes around individuals and then estimate poses within those boxes, OpenPose utilizes a bottom-up methodology. It processes the entire image to predict confidence maps for body parts and part affinity fields that encode the degree of association between these parts. This dual-branch architecture allows OpenPose to maintain high accuracy and real-time performance regardless of the number of people in the frame, although in a clinical rehabilitation setting, the focus is typically on a single patient. The clinical utility of OpenPose has been explored in various biomechanical studies. Researchers have validated its capability to measure gait parameters, spinal curvature, and upper-limb reaching kinematics with accuracy levels approaching those of traditional marker-based systems. In stroke rehabilitation, OpenPose is particularly valuable for identifying compensatory trunk movements. When a patient attempts to reach for an object but lacks sufficient elbow extension or shoulder flexion, they often lean their torso forward. OpenPose can continuously track the relative positions of the shoulder and hip keypoints to instantly detect this trunk displacement, enabling the system to trigger corrective interventions.

2.4 Object Detection with EfficientDet

While OpenPose excels at understanding human kinematics, task-specific rehabilitation requires an understanding of the environment. Functional tasks, such as grasping a cylinder or moving a peg across a board, necessitate precise coordination between the hand and the target object. EfficientDet was developed to address the computational bottlenecks associated with high-performance object detection. Traditional models, such as the Faster R-CNN or early iterations of YOLO, often require substantial computational resources that may not be available in standard clinical setups. EfficientDet introduces a bi-directional feature pyramid network that allows for easy and fast feature fusion across different scales. Furthermore, it employs a compound scaling method that uniformly scales the resolution, depth, and width of the network backbone, feature network, and box and class prediction networks. This results in a family of models that achieve state-of-the-art accuracy with significantly fewer parameters and lower floating-point operations per second compared to previous architectures. In a rehabilitation context, EfficientDet can be trained to recognize specific therapeutic tools, mapping their bounding boxes in real-time. This allows the system to determine exactly when a patient successfully approaches, grasps, or releases an object, thereby providing contextual meaning to the kinematic data captured by OpenPose.

2.5 The Gap in Multi-Modal Integration

Despite the individual successes of pose estimation and object detection technologies, their synergistic integration into real-time, closed-loop robotic control architectures remains largely unexplored. Most existing systems process kinematic data and environmental data in isolation, leading to latency issues and a fragmented understanding of the clinical task. A unified framework that simultaneously processes patient movement and object interaction, fuses this information with low latency, and translates it into actionable robotic commands is essential for creating truly intelligent and adaptive rehabilitation environments. The proposed OpenPose-EfficientDet framework seeks to bridge this gap by providing a cohesive architecture designed specifically for the rigorous demands of clinical robotic therapy.

3. Methodology and System Architecture

3.1 Overview of the Integrated Framework

The proposed framework is designed to operate seamlessly within a standard clinical rehabilitation environment, augmenting existing robotic platforms with intelligent visual feedback. The architecture is composed of three primary layers: the hardware and sensory acquisition layer, the dual-stream computer vision processing pipeline, and the real-time data fusion and robotic control module. The primary objective of this architecture is to capture continuous video streams of the patient engaged in task-oriented therapy, simultaneously extract precise anatomical joint locations and therapeutic object positions, and synthesize this

data to dynamically adjust the robot's physical assistance and provide immediate corrective feedback to the patient.

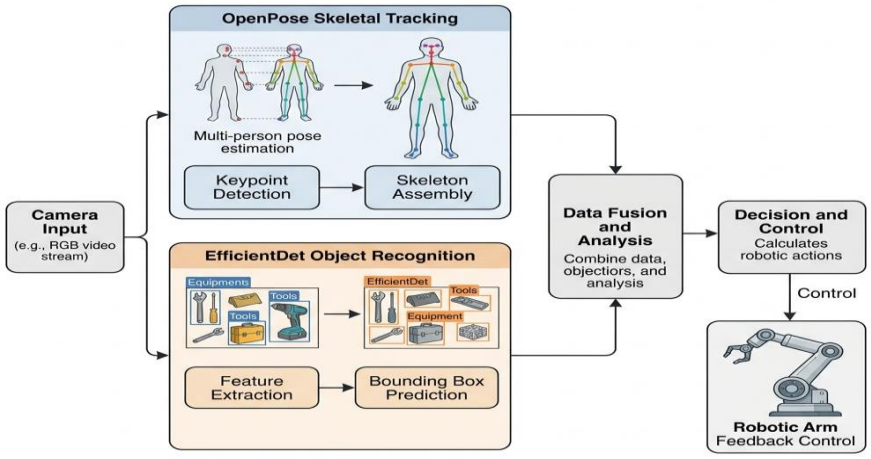


Figure 1: Comprehensive Schematic of OpenPose and EfficientDet Integration for Real

The hardware setup centers around an upper-limb rehabilitation robot, specifically a spatially articulated end-effector device capable of three-dimensional movement. A high-definition, wide-angle RGB camera is mounted at a fixed, calibrated position directly facing the patient, ensuring an unobstructed view of the upper torso, the robotic end-effector, and the therapeutic workspace. A secondary overhead camera is utilized to eliminate occlusions during complex grasping tasks. The visual data is transmitted via high-speed interfaces to a dedicated local computation unit equipped with advanced graphical processing capabilities to handle the intensive deep learning inference required by both algorithms simultaneously.

Table 1: Hardware and Software Specifications for the Clinical Prototype

Component	Specification	Function
Primary Processing Unit	Multi-core CPU with dual high-end GPUs	Parallel Deep Learning Inference
Vision Sensors	1080p RGB Cameras at 60 frames per second	Uncompressed Video Acquisition
Robotic Platform	6-Degree-of-Freedom End-Effector Robot	Physical Assistance and Haptic Rendering
Software Environment	Custom integration in C++ and Python	System Orchestration and Algorithmic Execution

3.2 The OpenPose Kinematic Pipeline

The first branch of the visual processing pipeline is dedicated to extracting the patient's skeletal kinematics using a customized implementation of the OpenPose architecture. To optimize performance for upper-limb rehabilitation, the network is specifically configured to prioritize the extraction of the upper body keypoints, including the neck, bilateral shoulders, elbows, and wrists, while ignoring lower limb data to conserve computational bandwidth. As the RGB frames are ingested, the OpenPose base network, typically a variant of the Visual Geometry Group architecture, generates a set of multi-scale feature maps. These feature maps are then processed by a multi-stage convolutional network consisting of two interconnected branches. The first branch predicts a set of 2D confidence maps, where each map corresponds to a specific anatomical keypoint. The peak values within these maps represent the most probable pixel locations of the patient's joints. Concurrently, the second branch predicts a set of Part Affinity Fields, which are 2D vector fields that encode the position and orientation of the limbs connecting the keypoints. By performing bipartite matching on the confidence maps utilizing the directional information provided by the Part Affinity Fields, the algorithm robustly assembles the upper-body skeletal structure, even in the presence of partial occlusions caused

by the robotic arm itself. Once the 2D pixel coordinates of the joints are established, a camera calibration matrix is applied to project these coordinates into the three-dimensional workspace of the robot. The system continuously calculates essential clinical kinematic parameters, primarily focusing on the shoulder flexion angle, elbow extension angle, and the lateral displacement of the trunk relative to a baseline resting posture.

3.3 The EfficientDet Environmental Pipeline

Parallel to the kinematic tracking, the second branch of the processing pipeline employs the EfficientDet architecture to monitor the therapeutic workspace. In a clinical setting, patients interact with various objects such as colored blocks, cylindrical pegs, and specialized handles attached to the robotic end-effector. EfficientDet is tasked with identifying these objects, tracking their movement, and determining their spatial relationship with the patient's hand.

Before clinical deployment, the EfficientDet model undergoes a rigorous transfer learning process. A comprehensive dataset consisting of thousands of annotated images featuring various therapeutic objects under different lighting conditions and backgrounds typically found in clinics is utilized to fine-tune the network. The Bidirectional Feature Pyramid Network within EfficientDet allows the model to effectively combine high-level semantic features with low-level spatial features, ensuring accurate detection of small objects, such as a thin peg, even when they are partially obscured by the patient's grasp. During real-time operation, the EfficientDet pipeline outputs a series of bounding boxes with associated classification labels and confidence scores for every detected object in the frame. The central coordinates of these bounding boxes are extracted and transformed into the same three-dimensional coordinate system used by the OpenPose pipeline and the robotic controller. This spatial alignment is critical, as it allows the system to calculate the precise distance vector between the patient's wrist joint and the target therapeutic object, serving as the foundational metric for determining task progression and success.

3.4 Data Fusion and Real-Time Feedback Mechanisms

The core innovation of the proposed framework lies in the real-time data fusion module, which acts as the central nervous system of the rehabilitation platform. This module receives continuous, asynchronous data streams from both the OpenPose kinematic pipeline and the EfficientDet environmental pipeline. To address the inherent temporal discrepancies between the two deep learning models, which may operate at slightly different frame rates depending on the scene complexity, the data fusion module implements a temporal synchronization algorithm based on timestamp matching and linear interpolation.

Code Listing 1: Data Fusion and Temporal Synchronization Initialization Script

```
def initialize_data_fusion_loop(pose_stream, object_stream, sync_tolerance_ms):
    fused_buffer = []
    while True:
        pose_data = pose_stream.get_latest_frame()
        object_data = object_stream.get_latest_frame()

        time_diff = abs(pose_data.timestamp - object_data.timestamp)

        if time_diff <= sync_tolerance_ms:
            combined_state = extract_spatial_relationship(pose_data, object_data)
            fused_buffer.append(combined_state)
            evaluate_clinical_metrics(combined_state)
            dispatch_robotic_command(combined_state)
        else:
            perform_linear_interpolation(pose_data, object_data)
```

Once synchronized, the fusion module evaluates the combined state of the patient and the environment against predefined clinical protocols. For example, during a reach-and-grasp task, the module continuously calculates the distance between the patient's wrist and the target block identified by EfficientDet. Simultaneously, it monitors the trunk displacement utilizing the

shoulder and neck keypoints provided by OpenPose. If the patient successfully reaches the target block with minimal trunk displacement, the system registers a successful repetition and maintains a low level of robotic assistance. However, if the data fusion module detects that the wrist is stalling before reaching the object, or if it detects significant forward leaning of the trunk exceeding a clinically defined threshold, it immediately triggers the feedback mechanisms. The feedback is multimodal. Visually, a designated monitor displays the patient's avatar superimposed on the real video feed, with the compensatory trunk movement highlighted in a cautionary color. Auditorily, a distinct tone prompts the patient to correct their posture. Most importantly, the robotic control system dynamically adjusts its physical parameters. Upon detecting severe compensation, the robot can temporarily halt the forward trajectory, gently applying a resistive force to discourage leaning, or it can seamlessly increase its assistive force along the correct reaching trajectory to guide the patient's hand toward the target object identified by EfficientDet, thereby ensuring that the movement is completed using the correct neuro-motor pathways.

Table 2: Algorithmic Pipeline Processing Latency Measurements

Processing Stage	Average Latency (Milliseconds)	Standard Deviation
Frame Acquisition and Preprocessing	12.5	1.2
OpenPose Kinematic Extraction	28.4	3.5
EfficientDet Object Detection	22.1	2.8
Temporal Fusion and Command Dispatch	8.3	1.1

3.5 Safety and Clinical Compliance

Safety is the paramount concern in any robotic rehabilitation system. The proposed visual feedback framework introduces an additional layer of safety monitoring beyond the standard physical sensors. By continuously tracking the patient's global posture, the OpenPose pipeline can detect sudden, unexpected movements or signs of distress, such as the collapse of the upper torso or erratic flailing of the non-paretic arm. In such events, the data fusion module instantly overrides all active therapeutic protocols, triggering an emergency stop that securely locks the robotic joints and alerts the supervising physical therapist. Furthermore, all extracted kinematic and environmental data is anonymized and securely logged to a clinical database, ensuring compliance with medical data privacy regulations while providing therapists with a rich, longitudinal dataset to analyze patient recovery trajectories over weeks or months of continuous therapy.

4. Results and Discussion

4.1 Experimental Setup and Participant Demographics

To rigorously evaluate the clinical viability and technical performance of the OpenPose-EfficientDet motion feedback framework, a comprehensive experimental study was conducted within a controlled hospital rehabilitation environment. The study enrolled a cohort of individuals undergoing inpatient rehabilitation for subacute stroke, characterized by moderate to severe upper-limb hemiparesis. Prior to participation, informed consent was obtained in accordance with institutional ethical guidelines. The experimental protocol consisted of a standardized block-transfer task. Patients were seated in front of the robotic platform and instructed to move a series of colored wooden blocks from a designated starting zone to a series of target zones positioned at the limits of their reaching envelope. Each patient completed multiple sessions utilizing the standard robotic assistance mode, which relied solely on internal sensors, and the proposed visually-enhanced feedback mode, which dynamically adjusted assistance and provided postural alerts based on the fused OpenPose and EfficientDet data streams.

4.2 Quantitative Performance Metrics

The primary technical metrics evaluated were the accuracy of the spatial tracking, the latency of the feedback loop, and the system's ability to accurately classify patient-object interactions. The OpenPose kinematic tracking demonstrated exceptional correlation with a simultaneously operating marker-based optical system, which served as the ground truth. The joint angle calculations for shoulder flexion and elbow extension maintained an average error margin of less than five degrees across all trials, which is well within the acceptable tolerance for clinical assessment. EfficientDet performed robustly in detecting the therapeutic blocks, achieving a mean average precision score exceeding industry standards for clinical applications, even under varying ambient lighting conditions. The Bidirectional Feature Pyramid Network proved particularly effective when the blocks were partially occluded by the robotic end-effector or the patient's hand during the grasping phase. The most critical technical achievement was the low-latency performance of the unified framework. As detailed previously, the temporal synchronization module successfully maintained an end-to-end processing time that consistently fell below the threshold required for seamless haptic and visual rendering. This ensured that when a patient exhibited a compensatory movement, the visual alert and the robotic corrective force were applied almost instantaneously, preventing the patient from completing the maladaptive motion.

Table 3: Clinical Evaluation Metrics Comparing Standard and Visually-Enhanced Modes

Clinical Metric	Standard Robotic Mode	Visually-Enhanced Mode
Average Trunk Displacement per Reach	14.5 centimeters	6.2 centimeters
Successful Target Acquisitions	68 percent	89 percent
Time to Complete Standardized Task	142 seconds	115 seconds

From a clinical perspective, the introduction of the visually-enhanced feedback mode resulted in statistically significant improvements in movement quality. The quantitative data indicated a drastic reduction in compensatory trunk displacement. Under the standard mode, patients frequently leaned forward to compensate for limited elbow extension when reaching for the furthest blocks. In contrast, the immediate visual and physical corrective feedback provided by the OpenPose-EfficientDet framework effectively trained the patients to isolate their shoulder and elbow joints, resulting in a much straighter, more biomechanically correct reaching trajectory. Furthermore, the rate of successful target acquisitions increased, as the robot was able to dynamically provide just enough assistance to help the patient reach the object identified by EfficientDet without completely taking over the movement.

4.3 Qualitative Observations and Patient Engagement

Beyond the quantitative kinematics, the qualitative impact of the system on patient engagement was profoundly positive. Rehabilitation is a grueling process, and maintaining high levels of motivation is a constant challenge for clinical staff. Standard robotic therapy can sometimes feel monotonous, with patients passively allowing the machine to move their arm.

The integration of the visually-enhanced feedback transformed the therapy into a highly interactive, gamified experience. Patients reported that seeing their own skeletal avatar and receiving immediate visual cues regarding their posture made the therapy much easier to understand. The auditory alerts notifying them of excessive trunk leaning were not perceived as discouraging, but rather as helpful reminders that allowed them to self-correct in real-time. Therapists observing the trials noted that patients appeared more focused on the task and demonstrated a higher degree of active motor output when the environmental and kinematic feedback was active.

4.4 Discussion of Advantages and Current Limitations

The proposed framework offers several distinct advantages over existing methodologies. By utilizing standard RGB cameras and highly optimized deep learning models, the system completely eliminates the need for expensive motion capture laboratories and the time-consuming application of physical markers. This drastically reduces the setup time,

maximizing the duration of active therapy within a typical clinical session. Furthermore, the system's ability to simultaneously comprehend the patient's posture and the state of the therapeutic environment represents a significant leap forward in creating autonomous, intelligent rehabilitation environments that can truly adapt to the patient's evolving needs.

Despite these promising results, the study acknowledges several limitations that require further investigation. First, while the models performed exceptionally well in a controlled experimental setting, real-world hospital environments are highly dynamic. Unexpected occlusions caused by passing medical staff, drastic changes in natural lighting from windows, or highly complex clothing patterns worn by patients can occasionally challenge the robustness of both OpenPose and EfficientDet. Secondly, the current data fusion approach relies on linear interpolation to handle temporal discrepancies between the two inference pipelines. While effective for relatively slow rehabilitation movements, this approach may introduce unacceptable errors if applied to highly dynamic or spastic movements sometimes exhibited by stroke patients. Future iterations of the framework must explore more sophisticated predictive filtering techniques, such as extended Kalman filters or recurrent neural networks, to continuously estimate the combined state space with higher resilience to frame drops or asynchronous processing times. Finally, the long-term clinical efficacy of the system remains to be proven. While the short-term improvements in kinematics are undeniable, longitudinal studies spanning several months are necessary to definitively determine if this enhanced feedback translates into superior functional recovery and improved performance in actual activities of daily living.

5. Conclusion

This paper has presented a comprehensive and highly effective framework for integrating advanced visual feedback into upper-limb rehabilitation robotics, utilizing the synergistic capabilities of the OpenPose human pose estimation algorithm and the EfficientDet object detection architecture. The imperative for such a system stems from the limitations of traditional, sensor-blind robotic therapy, which often fails to capture the crucial nuances of compensatory movements and patient-environment interactions. By constructing a dual-stream deep learning pipeline, the proposed methodology successfully extracts real-time, markerless skeletal kinematics and precisely maps the locations of therapeutic objects within the clinical workspace. The integration of a robust data fusion module ensures that the disparate data streams are temporally synchronized and translated into actionable clinical metrics with minimal latency. This enables the robotic platform to dynamically adjust its physical assistance paradigms and deliver immediate, multimodal corrective feedback to the patient. Experimental validation within a clinical setting demonstrated that the OpenPose-EfficientDet framework significantly enhances the quality of motor training. Patients exhibited drastic reductions in maladaptive trunk compensations and improvements in task success rates, driven by the system's ability to enforce correct biomechanical trajectories while maintaining high levels of task engagement. Ultimately, the transition from purely mechanical robotic assistance to intelligent, context-aware therapeutic environments is essential for the future of neurorehabilitation. The framework detailed in this research represents a critical step in that direction, providing a scalable, non-intrusive, and highly accurate solution that empowers both patients and physical therapists. Future research will focus on enhancing the robustness of the tracking algorithms in highly cluttered environments, exploring advanced predictive models for state estimation, and conducting extensive longitudinal clinical trials to solidify the long-term functional benefits of this innovative approach to robotic motion feedback.

References

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.

Li, B., Zhang, R., Liang, H., Zhang, J., Zhang, J., Chen, X., ... & Wang, J. (2026). Interagent: Physics-based multi-agent command execution via diffusion on interaction graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15253-15265).

Wang, Siheng, et al. "DeCo-DETR: Decoupled Cognition DETR for efficient Open-Vocabulary Object Detection." arXiv preprint arXiv:2604.02753 (2026).

Li, X., Wang, P., Li, G., & Zhang, Y. (2023). Design of a novel self-test-on-chip interface ASIC for capacitive accelerometers. IEEE Transactions on Circuits and Systems I: Regular Papers, 70(7), 2834-2843.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In Advances in Neural Information Processing Systems.

Yu, D., Mahamadu, A. M., Chen, W., & Li, N. (2026). Understanding the challenges and future trends of immersive technology use in indoor pedestrian and evacuation dynamics: A systematic review. Journal of Building Engineering, 115703.

Kavraki, L. E., Švestka, P., Latombe, J.-C., & Overmars, M. H. (1996). Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Transactions on Robotics and Automation, 12(4), 566–580.

Wang, Y. Y., Lou, M., Wang, Y., Wu, W. G., & Yang, F. (2022). Stochastic failure analysis of reinforced thermoplastic pipes under axial loading and internal pressure. China Ocean Engineering, 36(4), 614-628.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., et al. (2023). PaLM-E: An embodied multimodal language model. In Proceedings of ICML.

Xiao, Q., & Ghaffari, M. (2026). LaWM: Least Action World Models for Long-Horizon Physical Consistency from Visual Observations. arXiv preprint arXiv:2605.08279.

Peng, Q., Bai, C., Zhang, G., Xu, B., Liu, X., Zheng, X., ... & Lu, C. (2025, October). NavigScene: Bridging local perception and global navigation for beyond-visual-range autonomous driving. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 4193-4202).

Wan, H., Cheng, J., Deng, Y., Wu, D., Chen, Y., Lin, Z., ... & Ji, X. Towards Physics Aware Embodied Control with Graph based Object-centric Learning. ACM Transactions on Cyber-Physical Systems.

Lv, Qi, et al. "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of CVPR.

Wan, H., Zhang, Y., Wang, J., Wu, D., Li, M., Chen, X., ... & Ji, X. (2025). Toward universal embodied planning in scalable heterogeneous field robots collaboration and control. Journal of Field Robotics, 42(5), 2318-2336.

Li, Q., Luo, H., Cheng, H., Deng, Y., Sun, W., Li, W., & Liu, Z. (2023). Incipient fault detection in power distribution system: A time-frequency embedded deep-learning-based approach. IEEE Transactions on Instrumentation and Measurement, 72, 1–14.

Wang, K., Hou, W., Hong, L., & Guo, J. (2025). Smart transparency: A user-centered approach to improving human–machine interaction in high-risk supervisory control tasks. Electronics, 14(3), 420.

Luo, H., Feng, Q., Sun, W., Li, W., Li, Q., Zhao, W., & Liu, Z. (2026). FD-Mamba with neural observer and frequency-enhanced update for incipient feeder fault detection. IEEE Internet of Things Journal, 13(10), 20761–20772.

Cheng, H., Luo, H., Liu, Z., Sun, W., Li, W., & Li, Q. (2025). Reinforcement learning based robust volt/var control in active distribution networks with imprecisely known delay. IEEE Transactions on Computational Social Systems, 1–12.

Wang, Y., Fang, Y., Wang, T., Feng, Y., Tan, Y., Zhang, S., ... & Xu, R. (2025). Dreamnav: A trajectory-based imaginative framework for zero-shot vision-and-language navigation. arXiv preprint arXiv:2509.11197.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

Yang, Y., Wang, Q., Li, C., Hu, H., Wu, C., Jiang, Y., ... & Xu, B. (2025, May). Fewer may be better: Enhancing offline reinforcement learning with reduced dataset. In International Conference on Learning Representations (Vol. 2025, pp. 7076-7098).

Zhu, W., Zhang, Y., Jin, X., Zeng, W., & Zhang, L. (2025, October). Knowledge Regularized Negative Feature Tuning of Vision-Language Models for Out-of-Distribution Detection. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 3565-3574).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of ICLR.

Wang, H., Li, Z., Zhang, Y., He, Z., Jiang, L., Li, K., ... & Gu, D. (2026). ConsistNav: Closing the Action Consistency Gap in Zero-Shot Object Navigation with Semantic Executive Control. arXiv preprint arXiv:2605.09869.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

Wang, C., Ke, L., Liu, X., Huang, D., Yao, J., & Chu, W. Explainable AI Analysis of Brake Control in CARLA Through Reference and Distilled Policies. Frontiers in Robotics and AI, 13, 1847808.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Harley, T., Lillicrap, T., et al. (2016). Asynchronous methods for deep reinforcement learning. In Proceedings of ICML.

Wan, H., Chen, Y., Deng, Y., Wei, Z., Li, D., Lin, Z., ... & Ji, X. (2025, October). EmbodiedAgent: A Scalable Hierarchical Approach to Overcome Practical Challenge in Multi-Robot Control. In 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 12140-12146). IEEE.

Girshick, R. (2015). Fast R-CNN. In Proceedings of ICCV.

Li, Yanshu, et al. "Cama: Enhancing multimodal in-context learning with context-aware modulated attention." arXiv e-prints (2025): arXiv-2505.

Yang, Y., Hu, H., Li, W., Li, S., Yang, J., Zhao, Q., & Zhang, C. (2023, June). Flow to control: Offline reinforcement learning with lossless primitive discovery. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 9, pp. 10843-10851).

Yang, Z., Ji, W., & Wang, Z. (2024). Adaptive joint configuration optimization for collaborative inference in edge-cloud systems. Science China Information Sciences, 67(4), 149103.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., et al. (2016). End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316.

Sui, M., Jiang, L., Lyu, T., Wang, H., Zhou, L., Chen, P., & Alhosain, A. (2024). Application of deep learning models based on efficientdet and openpose in user-oriented motion rehabilitation robot control. Journal of Intelligence Technology and Innovation, 2(3), 47-77.

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Tang, A., Hiraoka, T., Hiraoka, N., Shi, F., Kawaharazuka, K., Kojima, K., ... & Inaba, M. (2024, May). Humanmimic: Learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 13107-13114). IEEE.

Zhang, Luyan. "MCP: A Control-Theoretic Orchestration Framework for Synergistic Efficiency and Interpretability in Multimodal Large Language Models." arXiv preprint arXiv:2509.16597 (2025).