This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: https://doi.org/10.1109/HPCC67675.2025.00106, 2025 IEEE International Conference on High Performance Computing and Communications (HPCC)

Wireless Single-Camera Markerless Motion Capture System for Healthcare Applications

Areej Athama
Department of Computer Science
Brunel University of London
United Kingdom
Areej.Athama@brunel.ac.uk

Apoorva Srivastava
Department of Computer Science
Brunel University of London
United Kingdom
Apoorva.Srivastava@brunel.ac.uk

Shengyang Huang
Department of Computer Science
and Electronic Engineering
University of Essex
United Kingdom sh24437@essex.ac.uk

Kezhi Wang
Department of Computer Science
Brunel University of London
United Kingdom
kezhi.wang@brunel.ac.uk

Yongmin Li
Department of Computer Science
Brunel University of London
United Kingdom
yongmin.li@brunel.ac.uk

Xiaojun Zhai
Department of Computer Science
and Electronic Engineering
University of Essex
United Kingdom xzhai@essex.ac.uk

***Abstract*—Single-camera markerless systems have emerged as a robust methodology for human motion capture and rehabilitation applications. Traditional methodologies typically necessitate multiple strategically positioned cameras or special equipment, including sensors to capture patient ambulatory motion, requiring preliminary calibration and synchronization procedures, which may incur significant costs. This paper presents a wireless single-camera markerless framework for rehabilitation applications that leverages advanced deep learning (DL) architectures to estimate and extract three-dimensional skeletal coordinates from monocular camera views of ambulatory patients. The extracted skeletal representation is subsequently transmitted across wireless communication channels. Then, the rendering technique has been applied for displaying virtual movement of the patient for privacy enhancement. Simulations demonstrate the effectiveness of the framework while maintaining motion assessment capabilities, presenting opportunities for deployment in remote healthcare monitoring scenarios.**

***Index Terms*—Wireless transmission, single-camera motion capture, markerless system, skeleton transmission**

I. Introduction

Markerless motion capture systems are gaining prominence as practical and cost-effective tools for gait analysis and movement assessment across diverse healthcare contexts, particularly for patients with neurological impairments and musculoskeletal disorders. These systems leverage recent advances in computer vision and deep learning (DL) to estimate human motion without requiring physical markers, offering the potential for remote and in-home clinical evaluations. Prior work, such as that by Scott et al. [1], has demonstrated the efficacy of single-camera markerless systems in capturing key gait and movement parameters within clinical settings. Despite their growing promise, these systems face limitations in accurately reconstructing detailed 3D skeleton information, an essential component for informed clinical decision-making [1]. Furthermore, while much of the existing literature focuses on the technical validation of these systems for improving markless motion capture, there remains a significant gap in understanding their practical integration and application in real-world clinical workflow [1], [2].

Recent studies have shown that, under controlled conditions, these systems can capture joint kinematics during overground walking with sufficient accuracy for clinical use [3]. However, their performance remains limited in scenarios that require precise 3D skeleton tracking, which can restrict their applicability for detailed or high-stakes clinical decision making [1]. In addition, it is not always possible to create a controlled condition for patients due to medical conditions or the inability to travel to the hospital.

Complementing the rise of markerless motion capture, remote assessment technologies that integrate video and wireless communication are becoming increasingly relevant in healthcare and rehabilitation due to their convenience, accessibility, and potential to support continuous monitoring. These systems enable clinicians to assess patients outside traditional clinical environments, reducing the burden of in-person visits. However, they also introduce technical and logistical challenges, such as high data transmission demands, particularly when dealing with volumetric or 3D video data, and increased concerns surrounding patient privacy. In this context, wireless single-camera markerless motion capture systems offer a promising middle ground. Their portability, ease of deployment, and cost-effectiveness make them well-suited for in-home assessments, especially for individuals with neurological or musculoskeletal conditions.

In this paper, we address the above challenges by presenting a comprehensive framework for wireless, near-real-time motion analysis using a single-camera, markerless approach. The proposed system extracts 3D skeletal data from 2D video inputs via a deep learning-based pose estimation model, i.e., PoseformerV2 [4], and transmits this information over a wireless communication channel. This enables usage of lesser bandwidth for transmission, as instead of the whole video containing subject and background information, only

crucial 3D skeleton data is transmitted. To enable efficient transmission, we employ a Deep Joint Source-Channel Coding (Deep-JSCC) model, similar to [5], which encodes the 3D skeletal data into high-dimensional embeddings using a multi-layer perceptron architecture, followed by decoding at the receiver end. The performance of the Deep-JSCC model is systematically evaluated through a series of experimental tests, demonstrating its viability for effective and nearly real-time markerless motion capture in wireless settings. In addition, we apply a privacy-preserving rendering framework that generates synthetic virtual human movement without the need for physical markers. This rendering pipeline integrates the SMPL models [6] and AMASS [7] to animate human figures based on 3D joint coordinates received from the decoder of the model. Crucially, the rendered videos emphasize skeletal movement while concealing identifiable attributes such as facial features and body texture, thereby preserving patient privacy and enabling secure remote consultations with healthcare providers.

The major contribution of this work includes:

- A wireless single-camera motion analysis framework is proposed that transmits only 3D skeletal data, reducing bandwidth usage.
- A privacy-preserving rendering block is introduced that reconstructs anonymized human motion videos from the decoded skeleton data, enabling secure and identity-free remote clinical assessments.

II. Background

A. Human Motion Applications

1) Markerless Motion Capture Applications: Markerless motion capture (MoCap) refers to techniques that track human movement without the use of physical markers, relying instead on video data and advanced algorithms to analyze motion. This technology has gained attention in clinical biomechanics, rehabilitation, and sports science due to its potential for cost-effective and accessible motion analysis [2]. It is also used in clinical settings for gait analysis and injury prevention. However, limitations such as reduced accuracy in 3D kinematics and dependency on environmental conditions (e.g., lighting) hinder their widespread adoption [2].

2) Single-Camera Markerless System: Recent advancements in computer vision and DL have enabled single-markless motion capture systems to develop, offering an accessible alternative to traditional multi-camera setups. These systems utilise single RGB cameras, such as smartphone cameras, to capture frontal-view videos of moving and walking subjects [8]. DL models can be used to extract 3D skeleton joints from these videos, providing accurate gait parameters without the need for complex calibration or synchronisation [9], [10]. Studies have demonstrated excellent validity and reliability of monocular systems compared to gold standard assessment tools, with correlations between classes ranging from 0.92 to 0.99 and percentage errors below clinical acceptability thresholds [10]. These systems show promise for widespread use in healthcare, sports, and rehabilitation settings, as they can be operated by non-technical personnel and provide comparable results to more complex and expensive multi-camera setups [8], [9].

3) Pose Estimation: Pose estimation is the process of determining the orientation and location of body segments in three dimensions. Although markerless systems can attain similar temporo-spatial measurements to marker-based systems, they may fail to precisely identify joint centre locations and angles, which are crucial for clinical applications [2]. Technological developments like the introduction of the MediaPipe [11], OpenPose [12], and Poseformer [4] have further made pose estimation easily available. However, they may lack discussion and applications or calibration in the medical and rehabilitation field.

B. Semantic Communication Techniques

Semantic communication (SC) has recently been widely discussed, focusing on conveying the meaning of information, rather than the exact bits or raw data that constitute the message [13]. SC may be implemented via several techniques, such as joint source-channel coding (JSCC) [14] and DL-based JSCC (Deep-JSCC) [5]. By emphasising the meaning of the data rather than just its bits, SC systems mark a paradigm shift in wireless data transmission, which may make transmitting large amounts of data efficient.

For wireless motion capture transmission systems enabled by SC, source coding can be used to compress video and extract skeletal data, while channel coding can improve resistance to noise. This approach may facilitate the evaluation of gait and posture through wireless communication channels, making video-based motion capture assessments both feasible and accessible, and potentially increasing the efficiency of healthcare services. Although SC systems show great promise in applications such as video transmission, their use in the healthcare and rehabilitation sectors remains limited [15].

III. Methodology

We introduce a framework, as shown in Fig. 1 that integrates a DL-based pose estimation model, i.e., PoseformerV2 [4] for extracting 3D skeleton from 2D camera video data, with the Deep-JSCC model for wireless data transmission, as well as SMPL models [6] and AMASS [7] for synthesizing anonymized human body animation from received skeletal data, thereby reducing wireless bandwidth requirements while maintaining the fidelity of motion information.

A. Problem Formulation

Let the 2D video be represented as $\boldsymbol{\mathcal{V}} = [I_1, I_2, \ldots, I_n, \ldots, I_N]$ where $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{F\times N\times M}$. Here F shows the number of frames in the video, and frames in the video are represented as $I_n \in \mathbb{R}^{N\times M}$ where $N\times M$ shows the spatial resolution of each frame. For all the frames present in the video $\boldsymbol{\mathcal{V}}$, the extracted 3D skeleton is represented as $\boldsymbol{\mathcal{J}}' = [J_1, J_2, \ldots, J_n, \ldots, J_N]$, $\boldsymbol{\mathcal{J}}' \in \mathbb{R}^{F\times K\times 3}$. The K means 3D joint coordinate extracted from each frame. In this study, $K = 17$ as 17 joints are extracted. These 17 joints are the hip, right hip, right knee,

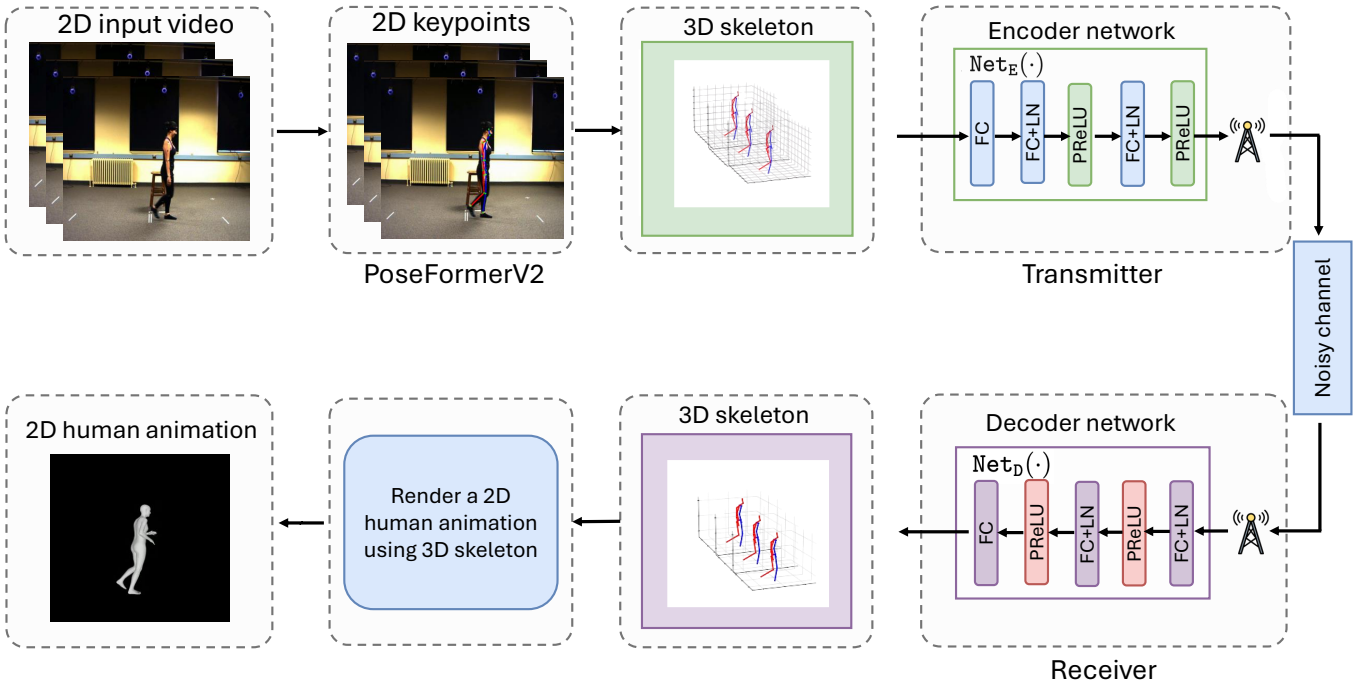

Fig. 1. The Wireless Single-Camera Markerless Motion Capture System

right foot, left hip, left knee, left foot, spine, thorax, neck, head, left shoulder, left elbow, left wrist, right shoulder, right elbow, and right wrist. For all the frames present in the video $\boldsymbol{\mathcal{V}}$, the extracted 3D skeleton is represented as $\boldsymbol{\mathcal{J}}' = [J_1, J_2, \ldots, J_n, \ldots, J_N]$, $\boldsymbol{\mathcal{J}}' \in \mathbb{R}^{F\times K\times 3}$.

The wireless transmission of $\boldsymbol{\mathcal{V}}$ is denoted by the function $\mathcal{T}(\cdot)$ and defined as,

$$\tilde{\boldsymbol{\mathcal{J}}} = \mathcal{T}(\boldsymbol{\mathcal{J}}') \tag{1}$$

where $\tilde{\boldsymbol{\mathcal{J}}}$ shows the reconstructed 3D skeleton at the receiver's end represented as $\tilde{\boldsymbol{\mathcal{J}}} = [\tilde{J}_1, \tilde{J}_2, \ldots, \tilde{J}_n, \ldots, \tilde{J}_N]$, $\boldsymbol{\mathcal{J}}' \in \mathbb{R}^{F\times K\times 3}$. Further, the received skeleton is input into the rendering block for the privacy preserving functions, denoted by the function $\mathcal{P}(\cdot)$ and defined as,

$$\hat{\boldsymbol{\mathcal{V}}} = \mathcal{P}(\tilde{\boldsymbol{\mathcal{J}}}) \tag{2}$$

where $\hat{\boldsymbol{\mathcal{V}}}$ shows the reconstructed virtual rendered moving objects at the receiver's end represented as $\hat{\boldsymbol{\mathcal{V}}} = [\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_n, \ldots, \hat{I}_N]$, $\boldsymbol{\mathcal{V}}' \in \mathbb{R}^{F\times N\times M}$. The ambition of this work is to display the virtual human motion video such that $\hat{\boldsymbol{\mathcal{V}}}$ is as close as possible with $\boldsymbol{\mathcal{V}}$ in terms of the moving angles of joints of patients for anonymous movement assessment in the rehabilitation applications.

B. 3D Skeleton Extraction

Aiming to understand the 3D human movement, it is important to extract the 3D human skeleton data from the video. In this work, the PoseformerV2 architecture [4] is used to extract the joints coordinate. The joint extraction process is represented as $\mathcal{E}(\cdot)$. For each frame, the 3D joint information is extracted by,

$$J_n = \mathcal{E}(I_n) \tag{3}$$

where $J_n \in \mathbb{R}^{K\times 3}$ shows the skeleton data from PoseformerV2.

C. Wireless Communication Model

1) System model: In this work, we aim to transmit the 3D skeleton coordinates stored in $\boldsymbol{\mathcal{J}}'$ via a wireless channel and reconstruct the skeleton at the receiver's end. The proposed model consists of three components: an encoder ($\texttt{Net}_\texttt{E}$), a noisy physical channel simulating the real environment, and a decoder ($\texttt{Net}_\texttt{D}$). Here, both the $\texttt{Net}_\texttt{E}$ and $\texttt{Net}_\texttt{D}$ are deep neural network models with trainable parameters. These trainable parameters are updated through back propagation. The $\texttt{Net}_\texttt{E}$ is defined as,

$$\mathbf{O} = \texttt{Net}_\texttt{E}(\boldsymbol{\mathcal{J}}', \alpha) \tag{4}$$

where α shows the SNR value of noise added in the encoder. $\mathbf{O} \in \mathbb{R}^{A\times C\times B}$. The $\boldsymbol{\mathcal{J}}'$ is encoded to a higher dimension to extract the important features from $\boldsymbol{\mathcal{J}}'$. During the transmission of the encoded bits to the receiver, noise is introduced by the physical communication channel. Here, noise such as additive white Gaussian noise (AWGN) is added as,

$$\tilde{\mathbf{O}} = \mathbf{O} + \beta \tag{5}$$

where $\beta \in \mathbb{R}^{A \times C \times B}$ denotes the complex normal Gaussian noise, which is $\beta \sim \mathcal{CN}(0, \sigma^2 I)$. The corrupted bits are passed through the decoder to generate the skeleton data, denoted as,

$$\tilde{\boldsymbol{\mathcal{J}}} = \mathtt{Net}_{\mathtt{D}}(\tilde{\mathbf{O}}, \alpha) \tag{6}$$

where $\tilde{\boldsymbol{\mathcal{J}}}$ is the reconstructed skeleton at the receiver's end after transmission.

2) Deep-JSCC based solution: The objective of the proposed model is to accurately reconstruct the original 3D skeletal data from encoded representations that may be corrupted due to transmission noise. The model comprises an encoder–decoder structure, where both $\mathtt{Net}_{\mathtt{E}}$ and $\mathtt{Net}_{\mathtt{D}}$ consist of fully connected layers with learnable nonlinearities and normalization mechanisms.

In $\mathtt{Net}_{\mathtt{E}}(\cdot)$, the input 3D joint coordinates are first projected to a hidden dimension using an FC layer with 64 neurons. This is followed by two additional FC layers with 128 and 256 neurons, respectively. Each linear transformation is followed by a Layer Normalization (LayerNorm) layer, which dynamically modulates the normalized output using a conditioning vector derived from the input. This adaptivity enables the network to normalize features in a context-aware manner, improving stability across varying input distributions. Nonlinearity is introduced via Parametric ReLU (PReLU) activations, which, owing to their learnable parameters, provide flexibility in shaping activations and mitigate the "dying neuron" issue.

During transmission, noise is introduced into the encoded features to simulate the real-world degradation, allowing the model to learn robust reconstructions.

At the decoder side, the architecture $\mathtt{Net}_{\mathtt{D}}(\cdot)$ decodes the noisy high-dimensional features back to the 3D joint space using a mirrored structure. The decoder first applies FC layers with 256, 128, and 64 neurons, successively reducing the dimensionality of the features. As with the encoder, each FC layer is followed by a LayerNorm layer and a PReLU activation. The final linear layer maps the decoded features back to the original 3D joint coordinate space.

D. Data preprocessing

For different subjects, the 3D skeleton coordinates extracted are based on the subject's position in space. Aiming to avoid the variation in the subject's coordinate, normalizing the coordinates is denoted as,

$$J'_{n,k} = \frac{J'_{n,k} - \frac{\sum_{n=1}^{N} J_{n,k}}{N}}{\sqrt{\frac{\sum_{n=1}^{N} (J_{n,k} - \mu_{J_k})^2}{N}}} \tag{7}$$

where $\mu_{J_k} \in \mathbb{R}^3$ denotes the average value of 3D joint, $J_{n,k} \in \mathbb{R}^3$ denotes the k-th 3D joint out of 17 joints from n-th frame, N denotes the length of all frames. After data preprocessing, the model transmits all 17 extracted 3D joints in the wireless channel.

E. Training details

The network is trained to minimize the mean square error loss, $\mathcal{L} = \mathbb{E}||\boldsymbol{\mathcal{J}}' - \tilde{\boldsymbol{\mathcal{J}}}||_2^2$. Here, $||\cdot||_2$ shows the L_2 norm and $\mathbb{E}$ represents the expectation function. The model is trained for 100 epochs with a batch size of 64. The SGD optimizer is used to find the optimum model.

F. Rendered 3D Skeleton framework

The design framework generates human body movement from 3D joint skeleton data by integrating large motion capture datasets and temporal pose estimation with parametric human body models. The framework integrates the rendering tools from AMASS [7] along with SMPL models [6] to establish a complete pipeline for human motion synthesis. SMPL model utilizes 24 axis-angle rotation vectors to represent the relative rotation angles of 24 human joints. These rotation vectors can be computed from transmitted 3D joint coordinates. Since the transmitted 3D skeleton only includes 17 joint coordinates compared to 24 rotation vectors in the SMPL model, the missing rotation vectors are set to zero. This approximation may impact our human body animation, but not too much, as the necessary human joints are included in the transmitted 3D skeleton.

With the rotation vectors, the SMPL model generates the new meshes that represent the current pose of the human body. Then the new mesh is used to render the human body using the rendering tool from AMASS. By changing the rotation vector of the pelvis, we can generate unlimited human body animation with different root orientations, which benefits to the doctors' diagnosis since they can see the walking actions of patients from any viewpoint.

IV. Experimental Details

The model is trained and tested on the publicly available datasets named **Large Multipurpose Motion and Video (MoVi)**. The datasets have been created on 90 subjects performing 20 predefined everyday actions and sports movements, using different hardware systems, including an optical motion capture system, video cameras, and inertial measurement units (IMU). This dataset contains motion capture data, video and IMU data [16].

A. Evaluation

Accurate extraction of joint angles is a critical aspect of human motion analysis, as these angles provide essential insights into biomechanical movement patterns. Therefore, in this study, we evaluate the performance of our model by comparing the computed joint angles against reference measurements. Here, we computed four joint angles, which are the left elbow, right elbow, left knee, and right knee. These joints are computed as,

$$\mathbf{a} = \mathbf{J}_1 - \mathbf{J}_2, \quad \mathbf{b} = \mathbf{J}_3 - \mathbf{J}_2 \tag{8}$$

$$\theta = \cos^{-1}\left(\frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|}\right) \tag{9}$$

TABLE I
JOINT ANGLE AND CORRESPONDING JOINT COORDINATES

Joint angle	$\mathbf{J}_1$	$\mathbf{J}_2$	$\mathbf{J}_3$
θ_{le}	Left wrist	Left elbow	Left shoulder
θ_{re}	Right wrist	Right elbow	Right shoulder
θ_{lk}	Left hip	Left knee	Left foot
θ_{rk}	Right hip	Right knee	Right foot

For different joints angle, left elbow (θ_{le}), right elbow (θ_{re}), left knee (θ_{lk}), and right knee (θ_{rk}), the $\mathbf{J}_1$, $\mathbf{J}_2$, and $\mathbf{J}_3$ vary which are shown in Table I.

The performance of the proposed model was evaluated along two principal dimensions: (1) the deviation between the predicted joint angles and the corresponding ground truth values, and (2) the temporal consistency of joint angle sequences, assessed by measuring the similarity between consecutive frames. To access these two metrics as used,

- **Dynamic Time Warping (DTW)**: DTW calculates a quantitative measure of the similarity between two time series of data [17]. A lower DTW distance indicates that the two sequences are more similar. It is evaluated as,

$$\text{DTW}(\theta_i, \hat{\theta}_j) = \text{Dis}(\theta_i, \hat{\theta}_j) \tag{10}$$

$$+ \min \begin{cases} \text{DTW}(\theta_{i-1}, \hat{\theta}_j), \\ 2 \times \text{DTW}(\theta_{i-1}, \hat{\theta}_{j-1}), \\ \text{DTW}(\theta_i, \hat{\theta}_{j-1}) \end{cases} \tag{11}$$

  where $i, j = \{0, 1, \ldots, s, \ldots, S\}$. S refers to the total number of subjects. $Dis(\cdot)$ shows the Euclidean distance between two points. θ and $\hat{\theta}$ shows the actual and reconstructed joint angles where $\theta = \{\theta_{le}, \theta_{lk}, \theta_{re}, \theta_{rk}\}$ and $\hat{\theta} = \{\hat{\theta}_{le}, \hat{\theta}_{lk}, \hat{\theta}_{re}, \hat{\theta}_{rk}\}$.
- **Mean Per Joint Angular Error (MPJAE)**: MPJAE quantifies the absolute difference between the estimated joint angles and the corresponding ground truth values, thereby providing an aggregate measure of the model's accuracy across all joints. It is computed as follows:

$$MPJAE = \frac{1}{S} \sum_{i=1}^{S} \|\theta i - \hat{\theta} i\|_1 \tag{12}$$

  where $i = \{0, 1, \ldots, s, \ldots, S\}$, S refers to the total number of subjects used to estimated joints, $\theta = \{\theta_{le}, \theta_{lk}, \theta_{re}, \theta_{rk}\}$ signifies the estimated angle of the joint that serves as the gold standard, and $\hat{\theta} = \{\hat{\theta}_{le}, \hat{\theta}_{lk}, \hat{\theta}_{re}, \hat{\theta}_{rk}\}$ denotes the reconstructed joint angle.
- **Mean Square Error (MSE)**: MSE finds the average squared difference between joint values received at the receiver's end and actual skeleton joint values. it is computed as follows:

$$MSE = \frac{1}{N} \sum_{n=1}^{N} (\tilde{J}_n - J_n)^2 \tag{13}$$

where $n = \{0, 1, \ldots, n, \ldots, N\}$, N refers to the total number of estimated joints and $N = 17$ as 17 joints are considered in this work.

V. RESULT AND DISCUSSION

In this section, we evaluate the performance of the proposed framework by comparing joint angle estimation metrics DTW and MPJAE, before and after data transmission. These metrics, commonly adopted in gait analysis, were applied to assess the fidelity of 3D joint coordinate estimations under varying SNR conditions. The comparison of the rendering part will be our future work.

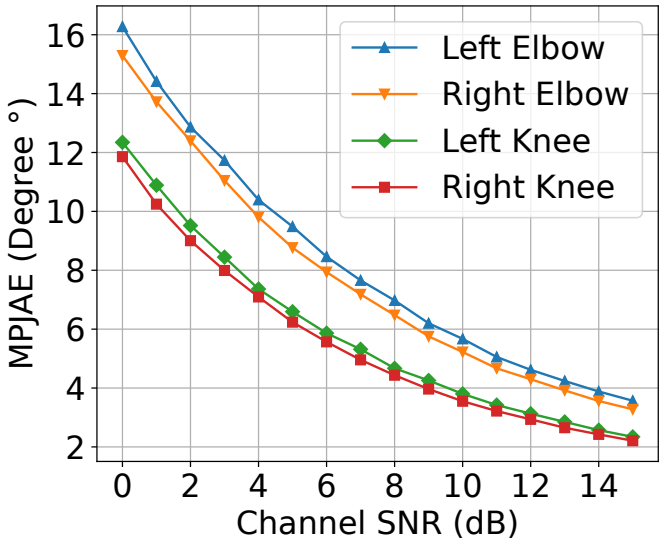

Fig. 2. MPJAE between the received skeleton and transmitted skeleton under different SNR.

The MPJAE and DTW values were obtained for AWGN channels, with SNR values ranging from 0 dB to 15 dB. Figure 2 and Figure 3 show the trend obtained for MPJAE and DTW with increasing SNR. MPJAE exhibited consistent trends, and decreases were observed with noise decreased (or SNR increased) for the right knee, right elbow, left knee and left elbow joints as follows: 81.36% (from 11.85 to 2.20), 81.03% (from 12.34 to 2.34), 78.57% (from 15.29 to 3.27), and 78.07% (from 16.27 to 3.56). Additionally, the DTW also showed the same trend of decrease for the right knee, right elbow, left knee, and left elbow joints as follows: 77.15%(from 5.27 to 1.20), 77.40% (from 5.54 to 1.25), 77.93% (from 7.38 to 1.62), and 77.50% (from 7.67 to 1.72). From the figure, one can see the impact of SNR on the performance of the wireless transmission.

Comparing it with the ground truth provided with the dataset, a similar trend was seen there as well. Figures 4 and 5 show the trend obtained for MPJAE and DTW with increasing

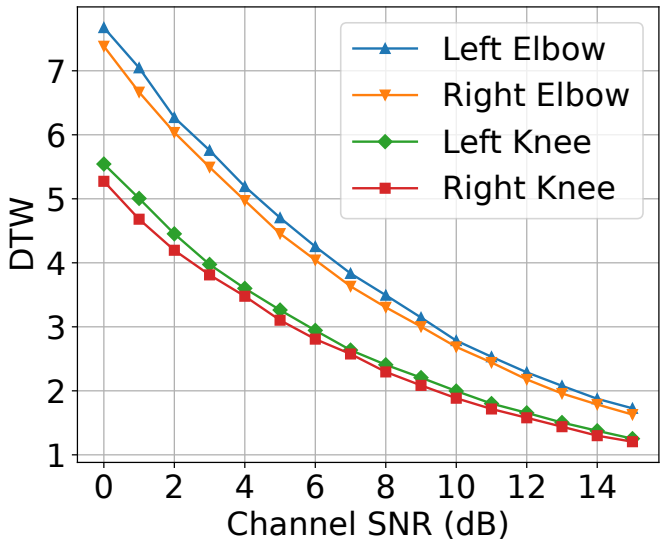

Fig. 3. DTW between the received skeleton and transmitted skeleton under different SNR.

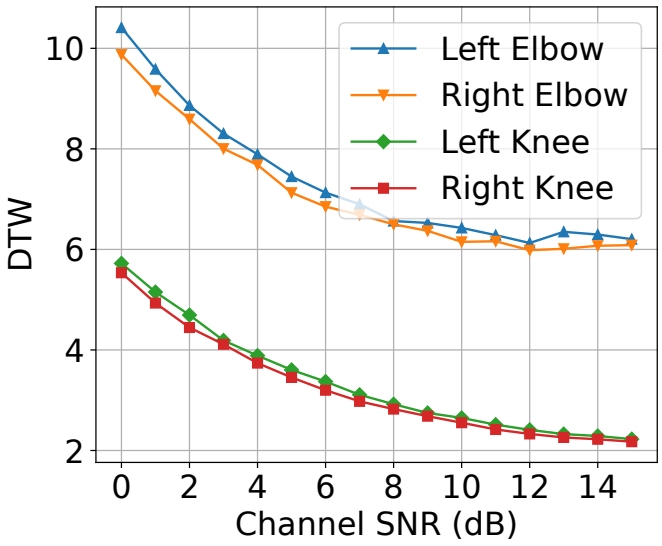

Fig. 5. DTW between the received skeleton and ground truth under different SNR.

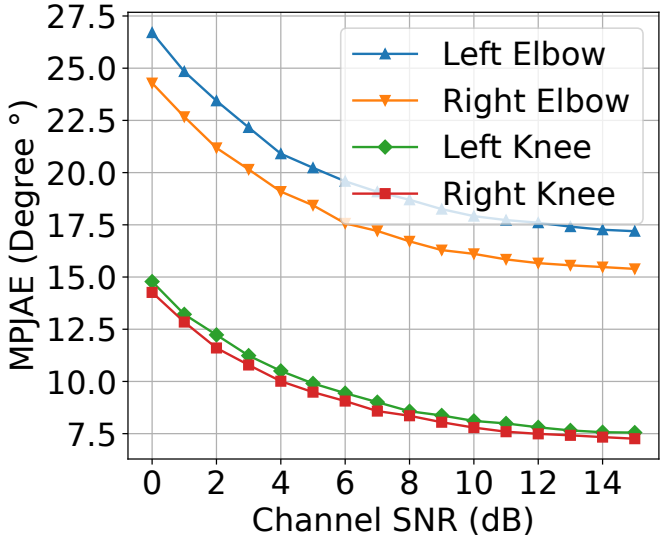

Fig. 4. MPJAE between the received skeleton and ground truth under different SNR.

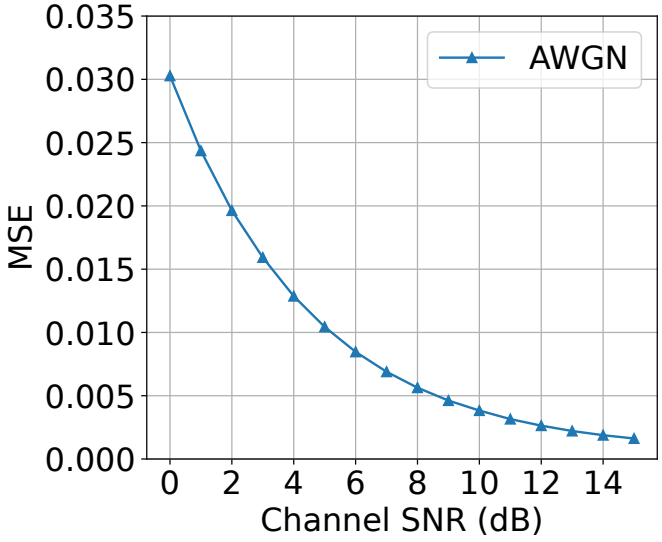

Fig. 6. MSE between the received skeleton and transmitted skeleton under different SNR.

SNR when compared with the ground truth. The decrease in MPJAE was observed for the right knee, right elbow, left knee and left elbow joints as follows: 49.11% (from 14.26 to 7.25), 48.91% (from 14.78 to 7.55), 36.63% (from 24.28 to 15.38), and 35.61% (from 26.70 to 17.19). Additionally, for DTW values, the decrease was 60.71% (from 5.53 to 2.17), 61.09% (from 5.72 to 2.22), 38.38% (from 9.87 to 6.08), and 40.41% (from 10.41 to 6.20) for right knee, right elbow, left knee and left elbow joints, respectively.

To evaluate the performance of the proposed model, the MSE was also calculated for all the joints across varying SNR levels. As evident from Figure 6, the MSE values decrease as the SNR values increases. On average, there is a 95% decrease in the values. Furthermore, to assess the fidelity of the reconstructed skeletons at the receiver's end, the MSE was also computed between the reconstructed data and the corresponding ground truth skeletons. As evident from Figure 7 there is 25% decrease in the MSE value as the SNR increases.

It is observed that the use of LN layer throughout the network serves to minimize internal covariate shift by conditioning the normalization process on the latent representation, leading to more stable and accelerated convergence during training. While PReLU activations promote sparsity, which

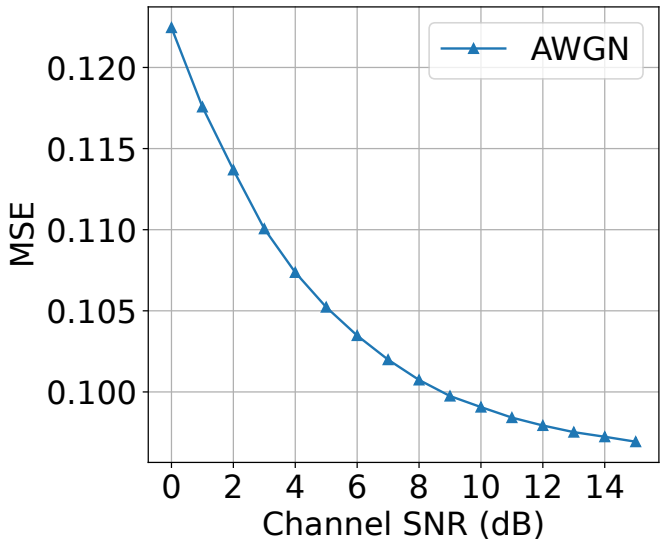

Fig. 7. MSE between the received skeleton and ground truth under different SNR.

is often desirable, caution is exercised in this context as excessive sparsity may impair the accurate recovery of fine-grained spatial information in the 3D skeleton data. These results highlight the robustness of the proposed model under various wireless transmission conditions, while also identifying the key areas, especially upper limb joint estimations, that are more susceptible to wireless channel noise. The proposed framework demonstrates broad applicability across biomechanics research, sports analysis, and entertainment industry applications, providing a robust foundation for future developments in human motion synthesis and analysis.

VI. Conclusion

In this paper, we present a framework for a single-camera, markless wireless motion capture system for healthcare and rehabilitation applications. The growing complexity of multi-camera systems underscores the need for a simpler solution. In contrast to the traditional system, our approach focuses on the transmission of skeletal keypoints, enabling efficient compression by isolating essential 3D joint information from background content, with the help of PoseformerV2 [4]. This strategy significantly reduces the volume of transmitted data while preserving critical motion features, particularly those relevant to health monitoring and analysis. After the transmitted 3D skeleton arrives at the receiver, we utilize the rendering tool from AMASS [7] and SMPL models [6] to render the human body animation using the 3D skeleton. The results demonstrate the effectiveness of the proposed model and indicate its strong potential for real-time deployment in practical applications. Our future work will be further investigating the usage of AMASS and SMPL for improved performance.

VII. Acknowledgment

We would like to acknowledge the support in part by the EU Eureka 5G4PHealth project (via Innovate UK, Grant No. 10093679), and the Horizon Europe HarmonicAI project Grant No. 101131117 (via UKRI Grant No. EP/Y03743X/1). K. Wang would like to thank the support from the Royal Society Industry Fellow scheme (IF\R2\23200104).

References

[1] B. Scott, M. Seyres, F. Philp, E. K. Chadwick, and D. Blana, "Healthcare applications of single camera markerless motion capture: a scoping review," *PeerJ*, vol. 10, pp. 13 517–13 544, 2022.

[2] L. Wade, L. Needham, P. McGuigan, and J. Bilzon, "Applications and limitations of current markerless motion capture methods for clinical gait biomechanics," *PeerJ*, vol. 10, pp. 12 995–13 022, 2022.

[3] M. Boldo, R. Di Marco, E. Martini, M. Nardon, M. Bertucco, and N. Bombieri, "On the reliability of single-camera markerless systems for overground gait monitoring," *Computers in biology and medicine*, vol. 171, pp. 108 101–108 113, 2024.

[4] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "Poseformerv2: Exploring frequency domain for efficient and robust 3D human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8877–8886.

[5] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[6] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *Seminal Graphics Papers: Pushing the Boundaries*, vol. 2, pp. 851–866, 2023.

[7] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Mass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, Jan. 2019, pp. 5442–5451.

[8] T. B. Rodrigues, D. P. Salgado, C. Ó. Catháin, N. O'Connor, and N. Murray, "Human gait assessment using a 3D marker-less multimodal motion capture system," *Multimedia Tools and Applications*, vol. 79, no. 3, pp. 2629–2651, 2020.

[9] X. Zhu, I. Boukhennoufa, B. Liew, C. Gao, W. Yu, K. D. McDonald-Maier, and X. Zhai, "Monocular 3D human pose markerless systems for gait assessment," *Bioengineering*, vol. 10, no. 6, pp. 653– 669, 2023.

[10] A. Azhand, S. Rabe, S. Müller, I. Sattler, and A. Heimann-Steinert, "Algorithm based on one monocular video delivers highly valid and reliable gait parameters," *Scientific Reports*, vol. 11, no. 1, pp. 14 065–14 075, 2021.

[11] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "MediaPipe: A framework for building perception pipelines," *ArXiv*, pp. 1–9, 2019.

[12] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[13] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE transactions on signal processing*, vol. 69, pp. 2663–2675, 2021.

[14] F. Zhai, Y. Eisenberg, and A. K. Katsaggelos, "Joint source-channel coding for video communications," pp. 1065–1082, 2005.

[15] A. Athama, K. Wang, X. Chen, and Y. Li, "Semantic communications for healthcare applications: Opportunities and challenges," in *2024 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, 2024, pp. 383–388.

[16] S. Ghorbani, K. Mahdaviani, A. Thaler, K. Kording, D. J. Cook, G. Blohm, and N. F. Troje, "MoVi: A large multi-purpose human motion and video dataset," *PloS one*, vol. 16, no. 6, pp. 0 253 157– 02 553 172, 2021.

[17] M. Müller, *Dynamic Time Warping*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.