International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2023)
J.-M. Normand, M. Sugimoto, and V. Sundstedt (Editors)

Stacked Dual Attention for Joint Dependency Awareness in Pose Reconstruction and Motion Prediction

L. Guinot[1], R. Matsumoto [1] and H. Iwata[2]

[1]Waseda University, Department of Modern Mechanical Engineering, Japan
[2]Waseda University, Faculty of Science and Engineering, Japan

Abstract

Human pose reconstruction and motion prediction in real-time environments have become pivotal areas of research, especially with the burgeoning applications in Virtual and Augmented Reality (VR/AR). This paper presents a novel deep neural network underpinned by a stacked dual attention mechanism, effectively leveraging data from just 6 Inertial Measurement Units (IMUs) to reconstruct human full body poses. While previous works have predominantly focused on image-based techniques, our approach, driven by the sparsity and versatility of sensors, taps into the potential of sensor-based motion data collection. Acknowledging the challenges posed by the under-constrained nature of IMU data and the inherent limitations in available open-source datasets, we innovatively transform motion capture data into an IMU-compatible format. Through a holistic understanding of joint dependencies and temporal dynamics, our method promises enhanced accuracy in motion prediction, even in uncontrolled environments typical of everyday scenarios. Benchmarking our model against prevailing methods, we underscore the superiority of our dual attention mechanism, setting a new benchmark for real-time motion prediction using minimalistic sensor arrangements.

CCS Concepts

• ***Computing methodologies*** → *Real-time simulation;* ***Motion processing;*** *Reconstruction;*

1. Introduction

The motion prediction and generation of avatars, particularly in Virtual and Augmented Reality, has grown increasingly paramount as technology's role in human-computer interaction continues to evolve. As VR and AR platforms continue to gain traction, their efficacy hinges on the realistic portrayal of user avatars, capturing the nuances of human motion. This requirement emphasizes the importance of generating realistic and kinematically consistent avatar motions.

Historically, generating realistic and kinematically consistent avatar motions was a challenging task, primarily due to the complexities associated with accurately predicting human movement. Traditional methods have often relied on extensive motion capture datasets, heuristic-based algorithms, or simplified kinematic models. While these approaches laid a foundational groundwork, they often fail to account for the nuance and unpredictability inherent in natural human motion and often stumbled in encapsulating the nuanced and spontaneous nature of genuine these motion. Such shortcomings become particularly stark within the domains of VR and AR, where immersion is contingent on authentic movement reproduction.

Recent advancements in deep learning and sensor technologies have opened doors to innovative approaches for motion prediction. Inertial Measurement Units (IMUs) – devices that measure velocity, orientation, and gravitational forces – have gained traction as essential tools for capturing and understanding human kinematics. Synthesised IMUs, in particular, offer a promising avenue by simulating the properties and benefits of physical IMUs without necessitating cumbersome equipment.

In this paper, we introduce a novel approach for pose reconstruction and motion prediction. Our method relies on the Natural Language Processing-inspired transformer's power of attention mechanisms, harnessing both spatial and temporal insights to create more accurate, fluid, and context-aware avatar movements. By seamlessly weaving spatial and temporal insights, our approach promises more precise, context-sensitive, and fluid avatar movements. The intrinsic dual attention framework, encompassing discrete modules for sensor-derived (spatial) and sequential attention, is architected to emphasize pivotal body joints and crucial time intervals. This focused lens ensures that generated motions resonate with human-like fluidity while upholding kinematic authenticity.

At the heart of our pose reconstruction and motion prediction strategy is a sophisticated attention mechanism. Our dual attention structure comprises:

- Sensor (Spatial) Attention:
  This module is dedicated to analyzing data from synthesized IMUs, emphasizing crucial body joints. By focusing on specific body parts, this spatial attention ensures the capture of intricate movements pivotal to human-like motion representation.
- Sequential Attention:
  Recognizing that movement is not just a factor of current physical state but also historical motion, our sequential attention module analyzes patterns over time. This aspect focuses on key time intervals, making it possible to predict subsequent movements based on historical data.

DOI: 10.2312/egve.20231326

Combining insights from both spatial and attention modules, our model generates realistic avatar movements. By considering both the immediate sensor data and the historical movement patterns, our approach ensures that the resulting motions are not just accurate but also context-aware, providing a fluid and natural avatar representation. Our methodology prioritizes real-time interaction and responsiveness. This emphasis ensures that while our approach is deeply rooted in rigorous research and theoretical kinematics, it translates effectively into practical avatar animation, suitable for immersive virtual experiences in VR/AR environments.

In essence, our method paints a comprehensive picture, amalgamating cutting-edge technology and deep understanding of human motion to offer a solution that promises both accuracy and practicality in the dynamic world of virtual representation. The overarching goal of our research is not just to achieve high-fidelity motion representation but also to facilitate real-time interaction and responsiveness in virtual environments. By bridging the gap between theoretical kinematics and practical avatar animation, our study offers a glimpse into the future of immersive virtual experiences.

2. Related Works

2.1. Image-based Motion Prediction

The image-based approach often dominates the motion prediction arena, primarily due to the vast availability of open-source datasets and certain constraints when relying solely on sensors. A prominent method within this category is motion capture, which, although widely adopted, is not without its limitations such as the need for multiple tracking markers, camera calibration, and specific background requirements. Addressing these issues, a marker-less strategy employing multiple cameras was proposed [N. 16]. While offline variants of this method demonstrated promising accuracy [BTG*12, BM98, HBL*17, J. 03], the real-time, online versions are usually favored in practice [ATS*08, EdJ*17, RRR*15, SHG*11]. However, challenges like camera calibration persist. Furthermore, Convolutional Neural Network (CNN) driven techniques employing a single stationary camera have been explored [CSWS17, CY14, HGDG17, NYD16, TJLB14, TS14]. Their primary limitation lies in producing results confined to a two-dimensional coordinate space.

In recent research, both offline [BKL*16, TMNSF17, ZZL*15] and online [MSM*18, MSS*17, OLPM*18] methodologies have been explored for the estimation of 3D posture from 2D images. In the context of human posture estimation, the focus has primarily been on predicting current postures. Notably, only two studies have ventured into the realm of future posture prediction. The first approach involves forecasting human motion 0.5 seconds ahead, utilizing detected human body joints from Kinect technology [HMS17]. The second approach, conducted using a single RGB camera, similarly predicts human motion after 0.5 seconds [WK19]. This second study leverages an Long Short-Term Memory network for capturing temporal information within images, incorporating Residual and Lattice Optical Flow to estimate subsequent postures and culminating in 3D reconstruction. It is worth noting that the domain of research dedicated to predicting future postures remains relatively under-explored, in stark contrast to the well-established field of human posture estimation. Furthermore, all of the discussed methods require a static camera viewpoint, demanding unobstructed visibility of the entire human figure. This limitation poses a challenge for predictions and estimations in scenarios with occluded image regions, where portions of the human subject may be obscured.

2.2. Sensor-based Motion Data Gathering

Contrary to image-centric methods, there are relatively fewer investigations into motion data techniques anchored entirely around sensor use. Particularly, there's a notable paucity of research focusing on harnessing sensor data for motion prediction. Current pose estimation studies utilizing inertial trackers (IMUs) and synthesizing accelerometer, gyroscope, and geomagnetic sensor data via a Kalman filter have been observed. Nevertheless, certain implementations, such as those described in [RLS09], rely on an extensive sensor count (up to 17), rendering sensor positioning and subsequent adjustments cumbersome. Efforts to diminish sensor counts have led to hybrid methods, intertwining sparse IMUs with video imagery [MGT*17, PBG*11, PBH*10, vMPMR16], optical markers [AHK*16], or depth cameras [HMST13]. Yet, these approaches often grapple with data loss during occlusions.

Human motion is inherently a complex kinematic chain [WAR17], with joint dynamics being interdependent. The collective movement of joints at any given moment t outlines the broader human movement trajectory. The position of a joint at one instance deeply impacts the subsequent poses. Recognizing this intricate interplay, our work posits the importance of perceiving these dynamics as time-series data, instrumental for future posture predictions.

A significant challenge when employing sensors lies in their partial information scope. With sensors offering data only at their specific placements, information gaps about intermediary body parts become glaring. For instance, having sensors only at the wrist and elbow leaves the arm's motion largely uncharted. This information scarcity amplifies with sparser sensor arrangements.

Considering these insights, this paper embarks on the mission of pioneering a motion prediction methodology. Rooted exclusively in IMU sensors, our approach aspires to holistically appreciate both the temporal context of human motion and the intricate sensor interdependencies.

3. Data and Environement

3.1. Synthetised data

Our method relies on a learning-based approach, which necessitates a significant dataset for training. While there are numerous datasets for camera or marker-based scenarios, there is a lack of public datasets that include both IMU data and precise poses. To our understanding, the sole dataset of this nature is TotalCapture [TGM*17], which captures standard daily activities. Given the scarcity of open-source datasets containing IMU sensor-based motion data, we devised a technique to translate motion capture data into an IMU-compatible format. In essence, each marker from motion capture yields data within a three-dimensional coordinate framework. Tapping into the Archive of Motion Capture as Surface Shapes (AMASS) repository [MGT*19], we harnessed motion capture data from comprehensive datasets like CMU [FdlTB08], HumanEva [SBB10], and JointLimit [AB15] to animate the SMPL model. The ability to match SMPL parameters with various data types (like IMUs, marker data) allows us to create a broader and more detailed training dataset. This is achieved by producing pairs of IMU readings and corresponding SMPL parameters from diverse datasets. To generate synthetic IMU training data, we put virtual sensors on the SMPL mesh. Using forward kinematics, we then gather orientation data, and determine accelerations through finite differences. The transformation to IMU data, encapsulating the rotation matrix, acceleration, and angular velocity, is achieved

through the instantiation of "virtual IMUs". These represent abstract renditions of sensors anchored onto the 3D SMPL structure, mimicking authentic IMU sensors in their data output characteristics in line with the SMPL model's movement dynamics. Strategically, these virtual IMUs were positioned at non-intrusive locations: right ankle, left ankle, waist, both wrists, and head. To synthesize these virtual sensors, we adhered to the following methodology:

- Acceleration Calculation:
  Acceleration, at its core, is deduced from the second derivative of positional variations. Given the motion capture's inherent frame rate of 120Hz, infinitesimal time intervals weren't feasible. As a remedy, the finite difference method (refer to Eq. 1) was employed where for a virtual IMU's position at time t is p_t and the gap between successive frames is dt, the simulated acceleration is calculated accordingly.

$$a_t = \frac{p_{t-1} + p_{t+1} - 2 \times p_t}{dt^2} \tag{1}$$

- Downsampling:
  To ensure compatibility with the minimal IMU sensor sampling rate (60Hz), the motion capture frames were downscaled by half prior to the computation.
- Rotation Matrix & Angular Velocity:
  Akin to the acceleration, the rotation matrix and angular velocity derivation adhered to a similar protocol. Angular velocity determination for each sensor was executed utilizing the finite difference methodology. Notably, previous studies did not incorporate angular velocity as a parameter, leaving us without a comparative benchmark for this computational approach. Let's denote the position of a sensor at a specific time instance t as R. The positional alteration between successive frames can thus be articulated as:

$$R_{diff} = R_{t-1}^{-1} R_{t+1} \tag{2}$$

  With a conversion between rotation matrix and axis angle $C_v(R)$: $RotationMatrix \mapsto AxisAngle$ expressed as:

$$\vec{\omega}_t = \frac{C_v(R_{diff})}{2dt} \tag{3}$$

This structured approach ensures a seamless conversion, bridging the gap between motion capture data and the requirements of IMU-based motion analysis.

3.2. Data

The methodologies embraced in this research are predominantly informed by paradigms in Natural Language Processing (NLP). Despite our primary dataset constituting time-series human motion information procured via IMU sensors, we opted for an NLP-centric data manipulation, particularly evident during the initial data formatting for network input.

For every discrete time instance denoted by t, data from each IMU sensor yields a 15-dimensional dataset. This encompasses a rotation matrix with 9 dimensions, an acceleration component of 3 dimensions, and a 3-dimensional angular velocity measure. Thus, this data can be conveniently abstracted as a 15-dimensional feature vector. Given our choice of employing 6 IMUs, the resultant output for each time instance becomes a 90-dimensional matrix, structured as "IMU count $\times$ Features" (Figure 1). Intriguingly, this matrix bears similarities to the "word embedding size" typically encountered in NLP contexts.

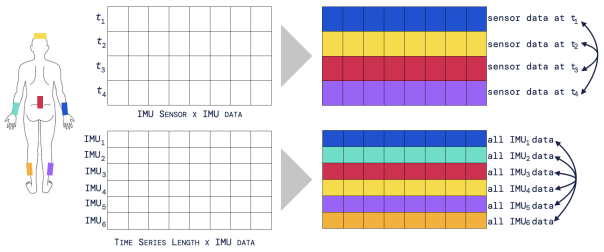

Figure 1: *Isolated IMU Attention (Bottom) and Sequential Attention (top)*

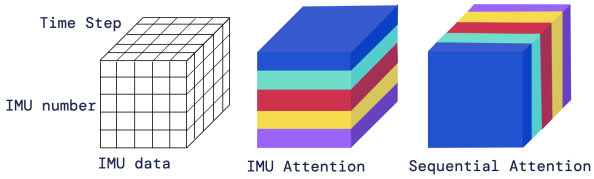

Figure 2: *3D representation of Dual Attention*

4. System Design

4.1. Sequential Attention (Self Attention)

Recurrent neural networks (RNNs) have proven adept at feature extraction from time-series data. Yet, over prolonged temporal spans, they periodically lose retrospective data and encounter growing challenges in parallel computations. In the realm of Natural Language Processing (NLP), the self-attention mechanism (as referenced in [VSP*17]) has emerged as a solution. It perceives word sequences within textual data as time-series information, emphasizing inter-word attention. Given a sentence, the mechanism discerns the significance of each constituent word. Crucially, this mitigates the issue inherent to extended time-series analysis, as it removes the predisposed notion that data at time instance t holds precedence over that at *t-1*. In the present project, self-attention was applied to rate the importance of all IMU data with respect to the time step. In this paper, this type of attention is referred to as "sequential attention"(Figure 1).

4.2. IMU Attention

IMU sensor-based posture prediction implies that data on an individual's current posture is confined to the regions adorned with sensors. Given that a posture transition arises from an amalgamation of diverse basic movements, as substantiated by [GTN18, FCTL18, GGM15], it becomes imperative to accurately depict this combination to faithfully represent the posture shift. Consequently, a holistic perspective that encompasses all sensors concurrently is deemed crucial, as opposed to an isolated examination of each sensor.

Introducing the notion of "IMU attention" we aspire to equip the system with a cognizance of inter-sensor relationships, emphasizing how motion dynamics at a particular point influence other regions.

4.3. Dual Attention

In this paper, Dual Attention can be seen as an extension of self-attention. Data used in the present paper is represented as the three dimensional matrix shown in Figure 2. Dual Attention refers to the

use of attention both in the "time-series direction" (sequential attention) and the "spatial direction" (IMU attention). The computational flow of this Dual Attention, comprised in what we call a "Dual Attention block" drawn in Figure 3 is as follows: First, two matrices, both for input tensors of 1 are generated:

$$I_{seq} \in \mathbb{R}^{Sequence \times (IMUNum \times Features)} \tag{4}$$

$$I_{IMU} \in \mathbb{R}^{IMUNum \times (Sequence \times Features)} \tag{5}$$

Query (Q), key (K) and value (V) vectors for sequential attention are defined as:

$$Q_{Seq} = I_{Seq} W_{Q_{Seq}} \tag{6}$$

$$K_{Seq} = I_{Seq} W_{K_{Seq}} \tag{7}$$

$$V_{Seq} = I_{Seq} W_{V_{Seq}} \tag{8}$$

With Q_{Seq},V_{Seq} and $K_{Seq} \in \mathbb{R}^{Sequence \times d_{Seq}}$ and $W_{Q_{Seq}}, W_{V_{Seq}}$ and $W_{K_{Seq}} \in \mathbb{R}^{(IMUNum \times Features) \times d_{Seq}}$. The W is a weight matrix in the form (d_{Seq}, Embedding size) with d_{Seq} an arbitrary parameter.

Similarly, for IMU attention:

$$Q_{IMU} = I_{IMU} W_{Q_{IMU}} \tag{9}$$

$$K_{IMU} = I_{IMU} W_{K_{IMU}} \tag{10}$$

$$V_{IMU} = I_{IMU} W_{V_{IMU}} \tag{11}$$

With Q_{IMU},V_{IMU} and $K_{IMU} \in \mathbb{R}^{IMUNum \times d_{IMU}}$ and $W_{Q_{IMU}}, W_{V_{IMU}}$ and $W_{K_{IMU}} \in \mathbb{R}^{(Sequence \times Features) \times d_{IMU}}$. Here again, W is a weight matrix in the form (d_{IMU}, Embedding size) with d_{IMU} an arbitrary parameter. In the present study, d_{Seq} and d_{IMU} were defined as:

$$d_{Seq} = IMUNum \times Features \tag{12}$$

$$d_{IMU} = Sequence \times Features \tag{13}$$

Using the query and key vectors, the attention ratio can be calculated, once the score has been determined. The later is done by using the internal product of the query and key vectors.

$$Score_{Seq} = Q_{Seq} K_{Seq}^T \in \mathbb{R}^{Sequence \times Sequence} \tag{14}$$

$$Score_{IMU} = Q_{IMU} K_{IMU}^T \in \mathbb{R}^{IMUNum \times IMUNum} \tag{15}$$

After normalizing these scores and with $\sqrt{d_{Seq}}$ and $\sqrt{d_{IMU}}$, the attention ratio (AR) is obtained from the Softmax function, applied respectively to each row of the matrix.

$$AR_{seq} = Softmax\left(\frac{Score_{Seq}}{\sqrt{d_{Seq}}}\right) \tag{16}$$

$$AR_{IMU} = Softmax\left(\frac{Score_{IMU}}{\sqrt{d_{IMU}}}\right) \tag{17}$$

The final output is a representation of the degree of relevance of each row element with respect to other rows - other time steps for sequential attention, other sensors for IMU attention.

$$Output_{Seq} = AR_{Seq} V_{Seq} \in \mathbb{R}^{Sequence \times d_{Seq}} \tag{18}$$

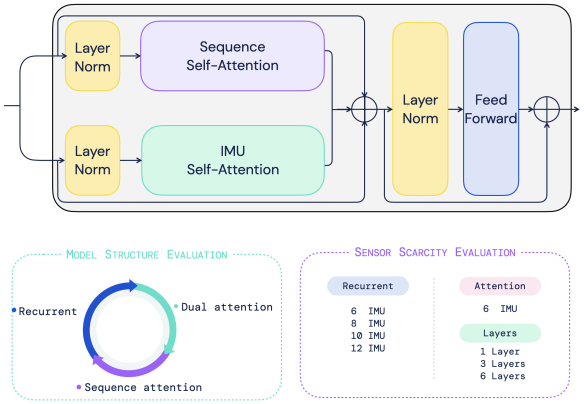

Figure 3: *Attention Block (top) and Evaluation Metrics (bottom)*

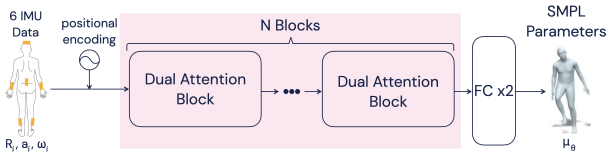

Figure 4: *Overall model image*

$$Output_{IMU} = AR_{IMU} V_{IMU} \in \mathbb{R}^{Sequence \times d_{IMU}} \tag{19}$$

Following these steps, it became possible to extract complex information about the relationship between each sensor data with respect to the time frame context, and about the spatial relationship of all sensors with respect to one another.

4.4. Overall model

As depicted in Figure 4, our comprehensive system takes rotation matrices, acceleration, and angular velocity data from virtual IMU sensors located at six distinct positions as its input. Positional encoding, is exclusively integrated into the sequential attention input. After processing through numerous multi-dimensional blocks (with N blocks being superposed) and culminating in a fully connected layer, the system outputs in the form of SMPL parameters, subsequently animating the respective figure.

The incorporation of positional encoding aims to maintain contextual integrity during the sequential information extraction. While the attention mechanism can effectively discern the interplay between time series and IMU data, it remains agnostic to the chronological sequence of time steps. In the context of IMU data and IMU attention, the precise order of the data might be non-essential as long as the core information is retrievable. However, when working with time series and sequential attention, any alteration in sequence can lead to a change in the data's inherent meaning. Consequently, to forestall any unintended contextual modifications, we applied positional encoding preceding the initial attention block.

The previous subsection, detailed how the attention ratio is calculated. However, additional operations need to be performed before data is input to the second layer normalisation (Figure 3). To avoid confusion, we will name the input of the overall attention blocks I'_{seq} and I'_{IMU}. First, residual connection is applied to the

output obtained by equations 18 and 19, to prevent gradient disappearance, a well known recurring issue in deeply connected layers.

$$Out\,put'_{seq} = Out\,put_{Seq} + I'_{seq} \in \mathbb{R}^{Sequence \times d_{Seq}} \tag{20}$$

$$Out\,put'_{IMU} = Out\,put_{IMU} + I'_{IMU} \in \mathbb{R}^{IMUNum \times d_{IMU}} \tag{21}$$

At the current stage, despite the successful calculation of sums, a discrepancy remains in the attention matrices' shapes between sequential and IMU attentions. By using d_{Seq} and d_{IMU}, the outputs were reshaped to:

$$\begin{aligned} &Out\,put'_{Seq} \in \mathbb{R}^{Sequence \times (IMU\,features \times features)} \\ &\rightarrow Out\,put''_{Seq} \in \mathbb{R}^{Sequence \times IMUNum \times features} \end{aligned} \tag{22}$$

and

$$\begin{aligned} &Out\,put'_{IMU} \in \mathbb{R}^{IMU \times (Sequence \times features)} \\ &\rightarrow Out\,put''_{Seq} \in \mathbb{R}^{Sequence \times IMUNum \times features} \end{aligned} \tag{23}$$

The Dual Attention mechanism output is as then defined by:

$$Out\,put_{Attention} = Out\,put''_{Seq} + Out\,put''_{IMU} \tag{24}$$

In the presented study, the loss is determined utilizing the Mean Squared Error (MSE) as described in Eq. 25. This loss is derived by contrasting the actual joint positions (y) of the SMPL model with the network's predicted joint positions ($\hat{y}$). Given the inherent history-based kinematic chain within the SMPL, the positional error of a specific joint is inherently influenced by its preceding joint's value. As a consequence, the frequency with which this error manifests varies between terminal joints (like the hand) and foundational joints (such as the waist).

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y})^2 \tag{25}$$

For the training process, we utilized the Adam optimizer with a starting learning rate of 1.0×10^{-3}. Building upon the insights of Kaichao et al. [YLJW19], which suggest that adaptive adjustment of the learning rate based on training progression can enhance learning efficiency, we employed a dynamic learning rate defined as:

$$learning\,rate = lr * \gamma^{epoch} \tag{26}$$

With lr the initial learning rate and γ is the attenuation factor (0.98). Training of the overall model was performed over an average of 80 epochs with early stopping.

5. Verification Experiment

5.1. Validation of the Proposed Method

- Benchmarking Dual-Attention mechanism:
  Our Dual Attention mechanism was critically benchmarked against prevailing methods, including a Bi-directional Recurrent Neural Network (BiRNN), as outlined in [HKA*18] and a sequential attention exclusive network (Figure 3, left). Comparative analyses were conducted with varying architectural complexities, specifically attention networks with 1, 3, and 6 layers.
- Effect of IMU Sensor Quantity vs. Attention Layers:

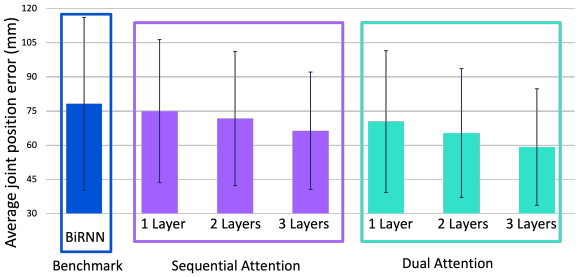

Figure 5: *Performance of different models*

An in-depth analysis was conducted to understand the performance variance resulting from altered IMU sensor counts as opposed to the inclusion of multi-dimensional attention layers (Figure 3, left).

5.2. System Performance Assesment

Synthesisted IMU sensor data, positioned on SMPL models, was the basis for our evaluation. The chosen BiRNN for this comparison stemmed from [HKA*18], owing to its conceptual and application-based congruence with our study. A self-attention exclusive network also served as a comparator. Performance metrics, beyond error rate, encompassed computational overhead and validation speed.

The data allocation strategy was split with 90% directed towards training and the residual 10% reserved for validation. Every set of 50 frames fed into the network projected pose estimations for the subsequent 30 frames. All experimental runs were executed in the Amazon Web Service P3 environment, leveraging eight Intel Xeon scalable VCPUs, cumulatively offering 61GB of CPU memory, and a NVIDIA Tesla V100 16GB GPU.

5.3. IMU Scarcity Analysis

While our model's foundational design envisaged the use of 6 sensors, we expanded the comparative horizon to 6, 8, 10, and 12 IMUs. Figure 8 illustrates the virtual positions (denoted by the coordinate system origins for each sensor). Adhering to the previously established methodology, the data partition was maintained at a 90:10 ratio for training and testing. Trials pivoted around the BiRNN model from [HKA*18], expanding the sensor count. Simultaneously, a Dual-Attention model with a steady sensor count (6), but escalating layer quantity, was used. The secondary experimental drive was to emphasize the superior capabilities of multi-dimensional attention. We aimed to assert that the elevated error rates observed in competing networks were not merely attributed to limited data capture points.

6. Results and Discussion

6.1. Model Structure Benchmarking

Figure 5 shows the prediction error results for all three tested neural networks. As depicted in Figure 5, a comparative error analysis was conducted across the three evaluated neural networks. It is noteworthy that models incorporating attention mechanisms, even those solely based on self-attention, outperformed the BiRNN model. For analogous sensor configurations and quantities, the integration of

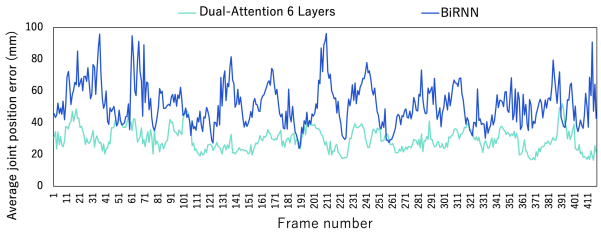

Figure 6: *Error rate comparison*

Figure 7: *Result example on static SMPL model*

even a singular sequential attention layer diminished the average joint position error by approximately 4 mm. Furthermore, introducing one layer of multi-dimensional attention yielded a reduction in average joint position error by an additional 10 mm, in contrast to the BiRNN results. For a more detailed perspective, Figure 6 presents a juxtaposition of joint position error between the BiRNN and the 6-layer Dual-Attention attention network during walking motion predictions.

Moreover, the efficacy of a singular multi-dimensional attention layer surpassed that of three layers focused solely on sequential attention. Insights from Table 1 accentuate that the Dual-Attention attention model not only yielded superior accuracy in terms of joint position error but also exhibited enhanced result generation frequencies relative to the sequential attention network. As a case in point, while a 6-layer sequential attention network registered an inference duration of 5.25 milliseconds, accompanied by an average error of 66.34 mm, a 3-layer Dual-Attention model showcased an average error of 35.37 mm within a swift 5.16 milliseconds inference span. Figure 7 shows a comparison of predicted posture with the ground truth in white and comparison models in green, red and blue. While discrepancies are visible on all three projections, the Dual-Attention model returned highest accuracy.

6.2. Sensor Scarcity

Prior research leveraging BiRNN networks for motion prediction deduced that amplifying the number of data capture nodes did not necessarily enhance result precision. This inference is further corroborated by our findings, as illustrated in Figure 8, which showcased a marginal reduction in joint position discrepancies. Nonetheless, the negligible variance in error metrics underscores that a mere augmentation of sensor count fails to rectify the inherent limitations of BiRNN. Such findings inevitably gravitate towards the assertion that BiRNN, in isolation, does not suffice for precise human posture extrapolation. It is worth noting that our analysis, constrained to 12 sensors, did not identify any ancillary studies hinting at considerable accuracy improvements with further sensor count escalations.

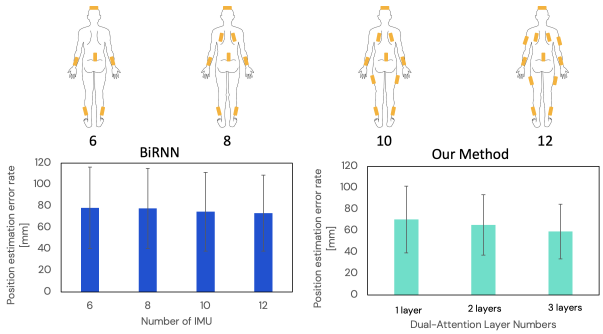

Figure 8: *Sensor scarcity and Attention layers relevancy*

Conversely, when maintaining a stable sensor count, each incremental addition to attention layers positively impacted the fidelity of ensuing pose predictions. Given our study's focal intent is on optimizing a minimalistic data acquisition system, we intentionally restrained the sensor count. However, considering the intrinsic objective of multi-dimensional attention—to discern inter-sensor relationships—it is plausible that incorporating a more extensive array of IMUs could further ameliorate the joint position estimation error.

6.3. Limitations

Prior investigations utilizing Bi-directional Recurrent Neural Networks (BiRNN) for motion prediction exhibited limited improvements in result accuracy when augmenting the number of data collection points. Surprisingly, our findings, as depicted in Figure 8, indicated a marginal reduction in joint position error with an increased sensor count. Nevertheless, this subtle decline in error rates failed to manifest a significant divergence, suggesting that a mere augmentation in sensor quantity does not effectively address the deficiencies of the BiRNN model. Consequently, our observations culminate in the assertion that a straightforward BiRNN framework inadequately supports precise human posture prediction. Notably, this study terminated its experimentation after evaluating performance with 12 sensors, as no extant literature was encountered to suggest substantial benefits in accuracy through further augmentation of IMUs.

Conversely, by maintaining a consistent sensor count, each increment in the number of attention layers yielded enhanced accuracy in future pose predictions. Given the overarching objective of this study to design a minimally data-intensive system, we deliberately constrained the sensor quantity. However, it is reasonable to extrapolate that employing an expanded array of IMUs could further diminish joint position estimation errors when leveraging the Dual-Attention mechanism [DCLT18, RWC*19]. This assumption aligns with the essence of Dual-attention, which is tailored to elucidate the intricate relationships interwoven among diverse sensors, thereby potentially facilitating the augmentation of IMUs to achieve heightened accuracy.

7. Conclusion

The domain of motion prediction and pose reconstruction has seen considerable evolution, with various techniques ranging from image-based methodologies to sensor-driven strategies. However, each approach comes with its own set of challenges and limitations,

Table 1: *Joint position error comparison*

	BiRNN	**Sequential**			**Dual Attention**		
layers		1	3	6	1	3	6
error (mm)	78.20	74.97	71.73	66.34	70.40	65.37	59.21
Standard Deviation	37.92	31.39	29.43	25.75	31.10	28.27	25.55
Inference (ms)	2.23	1.31	2.75	5.25	2.03	5.16	9.70

often requiring a compromise either in terms of accuracy, real-time applicability, or practicality.

In this paper, we introduced a pioneering method that harnesses a dual attention mechanism to reconstruct the human body pose in real-time using 6 Inertial Measurement Units (IMUs). This approach not only addresses the inherent under-constrained nature of pose parameters due to sparse IMUs but also circumvents the practical challenges posed by image-based methodologies.

By meticulously synthesizing insights from both image-based motion prediction studies and sensor-based research, our method offers a harmonious blend of accuracy, real-time performance, and user convenience. Our emphasis on the interdependencies among body joints and the temporal dynamics of human motion, as depicted in time-series data, offers a more comprehensive understanding of human movements. The novelty of using dual attention mechanisms – sequential and IMU attentions – provides a distinct edge in capturing these intricate dynamics over traditional methods.

While our work sets a promising precedent, it also opens doors for future research, especially in refining sensor placements, optimizing attention mechanisms for even sparser configurations, and potentially merging the strengths of image-based and sensor-driven approaches. As VR/AR technologies and other real-time applications continue to advance, the potential implications and applications of our work are vast, promising a more immersive and accurate user experience.

Our journey through this research has solidified our belief that while challenges in the realm of motion prediction are manifold, with innovative approaches and a keen understanding of the underlying mechanics, solutions are within reach.

References

[AB15] AKHTER I., BLACK M. J.: Pose-conditioned joint angle limits for 3d human pose reconstruction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1446–1455. 2

[AHK*16] ANDREWS S., HUERTA I., KOMURA T., SIGAL L., MITCHELL K.: Real-time physics-based motion capture with sparse sensors. pp. 1–10. 2

[ATS*08] AGUIAR E., THOBALT C., STOLL C., AHMED N., SEIDL H., THRUN S.: Performance capture from sparse multi-view video. *ACM Transactions on Graphics* (feb 2008). URL: 10.1145/1360612.1360697. 2

[BKL*16] BOGO F., KANAZAWA A., LASSNER C., GEHLER P., ROMERO J., BLACK M.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. vol. 9909, pp. 561–578. doi:10.1007/978-3-319-46454-1_34. 2

[BM98] BREGLER C., MALIK J.: Tracking people with twists and exponential maps. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (1998), pp. 8–15. 2

[BTG*12] BALLAN L., TANEJA A., GALL J., GOOL L. V., POLLEFEYS M.: Motion capture of hands in action using discriminative salient points. vol. 7577. 2

[CSWS17] CAO Z., SIMON T., WEI S.-E., SHEIKH Y.: Realtime multi-person 2d pose estimation using part affinity fields, 2017. 2

[CY14] CHEN X., YUILLE A.: Articulated pose estimation by a graphical model with image dependent pairwise relations. 2

[DCLT18] DEVLIN J., CHANG M.-W., LEE K., TOUTANOVA K.: Bert: Pre-training of deep bidirectional transformers for language understanding. 6

[EdJ*17] ELHAYEK A., DE AGUIAR E., JAIN A., THOMPSON J., PISHCHULIN L., ANDRILUKA M., BREGLER C., SCHIELE B., THEOBALT C.: Marconi—convnet-based marker-less motion capture in outdoor and indoor scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 3 (2017), 501–514. 2

[FCTL18] FANG H.-S., CAO J., TAI Y.-W., LU C.: *Pairwise Body-Part Attention for Recognizing Human-Object Interactions: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*. 09 2018, pp. 52–68. 3

[FdlTB08] FERNANDO DE LA TORRE JESSICA HODGINS A. B. X. M. J. M. A. C., BELTRAN P.: Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. In *Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University* (April 2008). 2

[GGM15] GKIOXARI G., GIRSHICK R., MALIK J.: Actions and attributes from wholes and parts. In *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 2470–2478. 3

[GTN18] GOUTSU Y., TAKANO W., NAKAMURA Y.: Classification of multi-class daily human motion using discriminative body parts and sentence descriptions. *International Journal of Computer Vision 126* (05 2018). 3

[HBL*17] HUANG Y., BOGO F., LASSNER C., KANAZAWA A., GEHLER P. V., ROMERO J., AKHTER I., BLACK M. J.: Towards accurate marker-less human shape and pose estimation over time. In *2017 International Conference on 3D Vision (3DV)* (Oct. 10–12, 2017), pp. 421–430. 2

[HGDG17] HE K., GKIOXARI G., DOLLÁR P., GIRSHICK R. B.: Mask R-CNN. *CoRR abs/1703.06870* (2017). URL: http://arxiv.org/abs/1703.06870. 2

[HKA*18] HUANG Y., KAUFMANN M., AKSAN E., BLACK M. J., HILLIGES O., PONS-MOLL G.: Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Trans. Graph. 37*, 6 (dec 2018). URL: https://doi.org/10.1145/3272127.3275108, doi:10.1145/3272127.3275108. 5

[HMS17] HORIUCHI Y., MAKINO Y., SHINODA H.: Computational foresight: Forecasting human body motion in real-time for reducing delays in interactive system. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces* (New York, NY, USA, 2017), ISS '17, Association for Computing Machinery, p. 312–317. URL: https://doi.org/10.1145/3132272.3135076, doi:10.1145/3132272.3135076. 2

[HMST13] HELTEN T., MÜLLER M., SEIDEL H., THEOBALT C.: Real-time body tracking with one depth camera and inertial sensors. In *2013 IEEE International Conference on Computer Vision* (2013), pp. 1105–1112. 2

[J. 03] J. STARCK AND A. HILTON: Model-based multiple view reconstruction of people. *Proceedings of the IEEE International Conference on Computer Vision*, 2 (nov 2003). URL: 10.1109/ICCV.2003.1238446. 2

[MGT*17] MALLESON C., GILBERT A., TRUMBLE M., COLLOMOSSE J., HILTON A., VOLINO M.: Real-time full-body motion capture from video and imus. In *2017 International Conference on 3D Vision (3DV)* (2017), pp. 449–457. 2

[MGT*19] MAHMOOD N., GHORBANI N., TROJE N. F., PONS-MOLL G., BLACK M. J.: AMASS: archive of motion capture as surface shapes. *CoRR abs/1904.03278* (2019). 2

[MSM*18] MEHTA D., SOTNYCHENKO O., MUELLER F., XU W., SRIDHAR S., PONS-MOLL G., THEOBALT C.: Single-shot multi-person 3d pose estimation from monocular rgb. pp. 120–130. doi:10.1109/3DV.2018.00024. 2

[MSS*17] MEHTA D., SRIDHAR S., SOTNYCHENKO O., RHODIN H., SHAFIEI M., SEIDEL H.-P., XU W., CASAS D., THEOBALT C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph. 36*, 4 (jul 2017). URL: https://doi.org/10.1145/3072959.3073596, doi:10.1145/3072959.3073596. 2

[N. 16] N. SARAFIANOS AND B. BOTEANU AND B. IONESCU AND I. A. KAKADIARIS: 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152 (2016). URL: https://www.sciencedirect.com/science/article/pii/S1077314216301369. 2

[NYD16] NEWELL A., YANG K., DENG J.: Stacked hourglass networks for human pose estimation. In *Computer Vision – ECCV 2016* (Cham, 2016), pp. 483–499. 2

[OLPM*18] OMRAN M., LASSNER C., PONS-MOLL G., GEHLER P., SCHIELE B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. pp. 484–494. doi:10.1109/3DV.2018.00062. 2

[PBG*11] PONS-MOLL G., BAAK A., GALL J., LEAL-TAIXÉ L., MÜLLER M., SEIDEL H., ROSENHAHN B.: Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *2011 International Conference on Computer Vision* (2011), pp. 1243–1250. 2

[PBH*10] PONS-MOLL G., BAAK A., HELTEN T., MÜLLER M., SEIDEL H., ROSENHAHN B.: Multisensor-fusion for 3d full-body human motion capture. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), pp. 663–670. 2

[RLS09] ROETENBERG D., LUINGE H., SLYCKE P.: Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. *Xsens Motion Technol. BV Tech. Rep. 3* (01 2009). 2

[RRR*15] RHODIN H., ROBERTINI N., RICHARDT C., SEIDEL H.-P., THEOBALT C.: A versatile scene model with differentiable visibility applied to generative pose estimation. In *Proceedings of the 2015 International Conference on Computer Vision (ICCV 2015)* (2015). 2

[RWC*19] RADFORD A., WU J., CHILD R., LUAN D., AMODEI D., SUTSKEVER I.: Language models are unsupervised multitask learners. 6

[SBB10] SIGAL L., BALAN A., BLACK M.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision 87* (03 2010), 4–27. 2

[SHG*11] STOLL C., HASLER N., GALL J., SEIDEL H.-P., THEOBALT C.: Fast articulated motion tracking using a sums of gaussians body model. pp. 951–958. 2

[TGM*17] TRUMBLE M., GILBERT A., MALLESON C., HILTON A., COLLOMOSSE J.: Total capture: 3d human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)* (2017). 2

[TJLB14] TOMPSON J., JAIN A., LECUN Y., BREGLER C.: Joint training of a convolutional network and a graphical model for human pose estimation. 2

[TMNSF17] TEKIN B., MÁRQUEZ-NEILA P., SALZMANN M., FUA P.: Learning to fuse 2d and 3d image cues for monocular body pose estimation, 2017. arXiv:1611.05708. 2

[TS14] TOSHEV A., SZEGEDY C.: Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1653–1660. 2

[vMPMR16] VON MARCAUD T., PONS-MOLL G., ROSENHAHN B.: Human pose estimation from video and imus. In *IEEE Trans. Pattern Anal. Mach. Intell.* (2016), vol. 38, pp. 1533–1547. 2

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A., KAISER L., POLOSUKHIN I.: Attention is all you need. 3

[WAR17] WANDT B., ACKERMANN H., ROSENHAHN B.: A kinematic chain space for monocular motion capture. 2

[WK19] WU E., KOIKE H.: Futurepose - mixed reality martial arts training using real-time 3d human pose forecasting with a rgb camera. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019), pp. 1384–1392. doi:10.1109/WACV.2019.00152. 2

[YLJW19] YOU K., LONG M., JORDAN M. I., WANG J.: Learning stages: Phenomenon, root cause, mechanism hypothesis, and implications. *CoRR abs/1908.01878* (2019). 5

[ZZL*15] ZHOU X., ZHU M., LEONARDOS S., DERPANIS K. G., DANIILIDIS K.: Sparseness meets deepness: 3d human pose estimation from monocular video. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 4966–4975. URL: https://api.semanticscholar.org/CorpusID:206594509. 2