EG UK Computer Graphics & Visual Computing (2024)
A. Slingsby and D. Hunter (Editors)

DeFT-Net: Dual-Window Extended Frequency Transformer for Rhythmic Motion Prediction

Adeyemi Ademola[1], David Sinclair[1], Babis Koniaris[1], Samantha Hannah[1], Kenny Mitchell[1]

[1] Edinburgh Napier University, UK

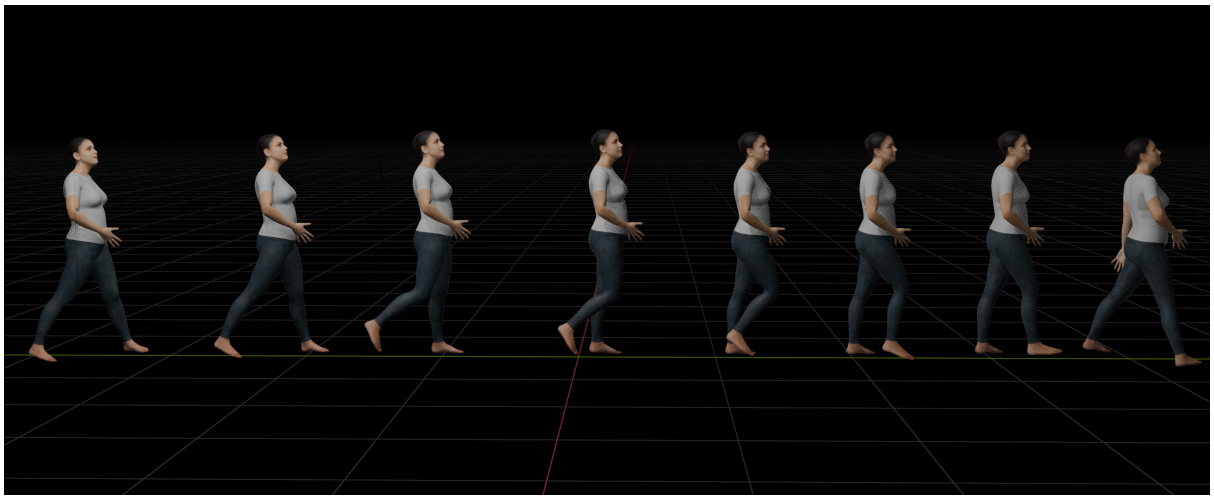

Figure 1: *An illustration of smooth plausible cyclic motion transition with improved prediction using our dual time window extended attention method of our DeFT-Net frequency transformer. Right foot anchor placement instance can be observed at the start of the transition through to end of the right foot's return to ground contact within the full walk cycle.*

Abstract

Enabling online virtual reality (VR) users to dance and move in a way that mirrors the real-world necessitates improvements in the accuracy of predicting human motion sequences paving way for an immersive and connected experience. However, the drawbacks of latency in networked motion tracking present a critical detriment in creating a sense of complete engagement, requiring prediction for online synchronization of remote motions. To address this challenge, we propose a novel approach that leverages a synthetically generated dataset based on supervised foot anchor placement timings of rhythmic motions to ensure periodicity resulting in reduced prediction error. Specifically, our model compromises a discrete cosine transform (DCT) to encode motion, refine high frequencies and smooth motion sequences and prevent jittery motions. We introduce a feed-forward attention mechanism to learn based on dual-window pairs of 3D key points pose histories to predict future motions. Quantitative and qualitative experiments validating on the Human3.6m dataset result in observed improvements in the MPJPE evaluation metrics protocol compared with prior state-of-the-art.

CCS Concepts

• ***Computing methodologies*** → *Machine Learning; Motion Processing; Virtual Reality;*

1. Introduction

In the fields of virtual reality (VR) and computer vision, real-time tracking is crucial for recovering accurate 3D pose data. Human

DOI: 10.2312/cgvc.20241220

joint pose data is commonly captured using multi-camera or single-camera setups integrated with AI algorithms to obtain depth information and directly recover pose key points and joint orientations. Nevertheless, challenges such as limited sensor range, occlusion, and latency persist in tracking 3D pose data. In order to improve immersion and engagement in patterned motion scenarios, there is a high demand for techniques that minimize latency [KSM24] [SAKM23] during motion tracking through motion prediction.

Deep learning techniques have significantly advanced the domain of human motion prediction [BBKK17] [CSY20]. Among these, recurrent neural networks (RNNs) have become particularly popular for predicting sequential human pose data [JZSS16] [FLFM15]. However, when it comes to long-term horizons and periodic motions, RNNs often struggle due to their inability to effectively capture long-term history, which is essential for forecasting periodic motion actions. To address this limitation, recent approaches have incorporated encoders [LZLL18] to better represent historical information.

Our work introduces a dual-window extended frequency attention-based human motion prediction technique that utilizes synthetically generated periodic data based on re-timed foot anchor placements, as illustrated in Fig. 2. Our method is motivated by the observation that humans tend to repeat their motions in actions such as dancing to music beats. To validate this, we focus on the context of rhythmic motion prediction, where we demonstrate the effectiveness of our approach by re-timing *Human3.6m* [IPOS13] to match these rhythmic patterns. We present results based on analyzing relevant information from significant bones, such as the feet, over a fixed-length period.

Inspired by previous works [MLSL19], we represent each subsequence of foot anchors in the trajectory space using a Discrete Cosine Transform (DCT). We then introduce our dual-windowed extended frequency motion attention as weights for DCT-encoded motion aggregation into a future motion estimate. To encode spatial dependencies between joints, we combine the motion estimate with the last observed matching period, using the result as input to a graph convolutional network (GCN) [FYD*23]. Our experiment, as shown in Fig. 4, demonstrates that our approach outperforms state-of-the-art methods in long-term and short-term periodic motion prediction on the Human3.6M walking and walking together datasets. Our work extends [MLS20], specifically improving 3D pose motion prediction for known periodicity based on foot anchor placements.

Our main contributions are summarized as follows:

- We analyze the causes of high errors in motion prediction and synthesize re-timed motion with supervised foot anchor information of periodic cycles, such as walking, for the defined use case of rhythmic motion prediction.
- We achieve superior overall mean per joint position error (MPJPE) results compared to state-of-the-art methods in experiments on the Human3.6M dataset for forecasting short and long-term motions by introducing *OurDualWindowDCT* attention aligned on a best fit major period of each motion sequence.
- We release an extension of the Human3.6M dataset with observed cyclic foot placement periods `https://github.com/CarouselDancing/DeFT-net`

2. Related Work

2.1. Traditional Approaches

Empowered by the probabilistic nature of the task of periodic dance motion, earlier methods like Boltzman and Hidden Markov Models [BH00] [THR06] have been employed to predict motion sequences. Their learned model synthesizes motion data via style interpolation and can be driven by 2D video, or script to generate new choreography for virtual motion-capture style synthesis. Despite the effectiveness of this procedure, it lacks accuracy and adaptability, particularly for capturing short and long dependencies for dynamic contexts like walking and dance sequences.

2.2. Recurrent Model Approaches

Over the years, RNNs have gained popularity in the task of 3D-human motion prediction [CAW*19]. An encoder-recurrent-decoder model (ERD) was proposed by *Fragkiadaki et al* [FLFM15], where a long short-term memory cell (LSTM) cell operates in a latent space. The work by *Jain et. al* [JZSS16] adopts an st-graph skeleton, applying the RNNs as nodes. The work by [AKH19] replaces dense output layers in the RNN architecture with structural prediction layers. This technique explicitly models joint dependencies that follow a kinematic chain. To accurately refine noisy RNN predictions, the work by [GSAH17] explicitly train a separate de-noising auto-encoder. All these methods suffer from the inability to capture long-range motion history.

To address the *transition problem* between a seed and prediction, *Martinez et. al* [MBR17] introduced a sequence-to-sequence (seq2seq) architecture with an input-to-output skip connection. To alleviate the exposure to bias problem , they also proposed training the model with the predictions. While this method led to better performance than previous pose-based works [FLFM15] [JZSS16], predictions still suffered from discontinuities between observed frames and predicted frames. Similary, a *teacher-forcing* ratio approach was adapted by [PGA18] to expose the model to its own predictions. Employing a hierarchy of RNNs, the work by [CAW*19] presented a seq2seq approach that can be modified to explicitly model different time scales. *Gui et al.* proposed adversarial training to generate smooth sequences [GWLM18]. *Ruiz et. al* approached human motion forecasting as a tensor impainting problem and adapted generative adversarial networks for long-term prediction in this work [HGMN19]. While this approach improves performance, the use of an adversarial classifier complicates training making it a challenge to deploy on periodic dataset with foot anchor encoding.

2.3. Beyond Recurrent Models

Given the drawbacks of RNNs, several works have employed the use of feed-forward networks as an alternative solution. [BBKK17] [MLSL19] The work by *Butepage et. al* [BBKK17] introduced a fully connected feed forward to process the recent history poses, investigating techniques to encode temporal historical information via convolution and exploiting the kinematic tree to encode spatial information. *Li et. al* suggests a convolutional sequence-to sequence model(CNN) processing a two-dimensional pose matrix

whose column represent the pose at every time step [LZLL18]. The model was employed to extract a pose motion prior from long-term motion history of frames, which, in conjunction with more recent motion history, was used as an input to an auto regressive network for future pose prediction. While more effective than RNN-based frameworks, the manually selected size of the convolutional windows highly influences the temporal encoding of motion sequences. To address this, *Aksan et. al* [AKCH21] introduced a spati-temporal transformer encompassing a fully auto-regressive approach to model temporal dependencies given the recursive nature of human motion. *Cai et. al* [CHW*20] leverage a transformer architecture on the DCT coefficients extracted from the seed sequence and make joint predictions progressively by following a kinematic tree. Similarly, *Mao et. al* [MLSL19] encodes joint sequence via DCT and train a graph convolutional network (GCN) to capture/learn inter-joint dependencies. Since the GCN operates on temporal windows of poses to produce an output, the pose forecast are limited to a predetermined length. To address this, [MLSL19] extracted DCT coefficients from shorter sub-sequences in a sliding window fashion aggregated with a 1D attention block. [GMI23] introduced a stacked-attention mechanism utilizing synthetic IMU data to improve long-term dependency handling in dance motion prediction. This method addresses the limitations of traditional RNNs by transforming motion dynamics into the frequency domain using discrete cosine transform (DCT), which better encodes temporal information.

Our work is related to these approaches, but differs in two aspects. First, windowed inputs we introduce a time beat signal based on foot anchor pose information to the DCT windowed input so our model can learn periodic motions of short and long term history in the frequency domain. We then introduce a dual-window extended frequency model to pay attention to periodic motions.

2.4. 3D-based Human Motion Capture Datasets

Human3.6M represents a significant advancement in human pose estimation by providing a large-scale dataset of 3.6 million accurate 3D human pose motions introduced by *Ionescu et. al* [IPOS13]. This dataset, much larger than previous ones, was created by recording 11 subjects (5 female and 6 male) from four different viewpoints, covering a wide range of typical human activities such as taking photos, talking on the phone, and eating. It includes synchronized images, motion capture, and depth data, along with accurate 3D body scans of the subjects. A unique feature of this dataset is its controlled mixed reality scenarios, allowing the study of human models under various conditions, including camera movement and occlusion. The dataset also includes extensive statistical models and detailed evaluation baselines, demonstrating its diversity and potential for future research.

AMASS stands as a groundbreaking development in the field of computer vision and motion analysis. In contrast to existing motion capture (mocap) datasets, which are often limited in size and scope, introduced by *Mahmood et. al* AMASS amalgamates 15 different optical marker-based mocap datasets into a singular, extensive collection. [MGT*19]. This is achieved through a novel method, MoSh++, which converts mocap data into detailed and realistic 3D human meshes. These meshes are represented by the SMPL model, a widely recognized framework known for its standard skeletal representation and fully rigged surface mesh, ensuring consistency and wide applicability. Mosh [LMB14] is versatile, accommodating various marker sets and capturing intricate details like soft-tissue dynamics and hand motions. Its accuracy and performance are fine-tuned using a new dataset of 4D body scans, recorded in conjunction with marker-based mocap. AMASS sets a new standard for human motion datasets with its rich compilation of over 40 hours of motion data, encompassing more than 300 subjects and over 11,000 motions.

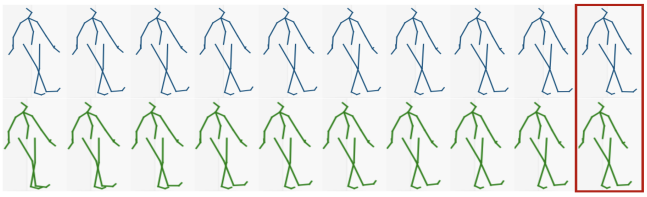

Figure 2: *A skeleton-grid comparison of the fixed DCT motions from the HistRepeatDCT method [MLS20] and our re-timed dual window extended DCT motions for test subject 5 walking together synchronised with right foot anchor placements. The fixed DCT motion sequence is shown as blue, and our dual window extended DCT motions as green skeleton. Note that the foot placement for our re-timed motions match exactly with fixed DCT motions at the last foot anchor frame.*

Motorica. Perez et al. [ANBH23] [VPHB*21] introduced a rich and diverse compilation of motion capture and audio recordings dedicated to various dance styles, totaling 6 hours. The data is standardized in BVH format and aligned to a single skeleton structure for consistency [VPHB*21]. A high-tech optical marker-based system collected it at 120 frames per second. The dataset covers four sessions, each with a unique style focus. The first session captures intricate street dance styles, emphasizing detailed finger movements. The second session shifts to casual dancing, set to a backdrop of pop music, though it omits the detailed finger motion capture found in other parts. In the third session, the spotlight is on vintage jazz dances, with an attempt to capture finger motions using specialized gloves, albeit with some quality issues due to sensor drift. The final session blends street dance styles with jazz, employing a simplified approach to finger motion capture. Overall, this dataset stands out for its comprehensive range of dance styles and meticulous approach to capturing motion, especially the varying levels of finger motion detail, making it an invaluable resource for digital dance movement analysis and research practice [JZSS16].

For the purposes of our experiment and validation, we select walking and walking together actions for 11 subjects in the Human3.6m dataset.

3. Method Overview

Our work introduces a unique approach to improving human motion prediction by incorporating periodic patterns and adapting a dual window of poses Z_i. Each Z_i consists of two concatenated slices S_i and $S_{i+p+\text{offset}}$ from the motion history $S_{1:N} = [s_1, s_2, s_3, \ldots, s_N]$, where p represents the period and offset allows

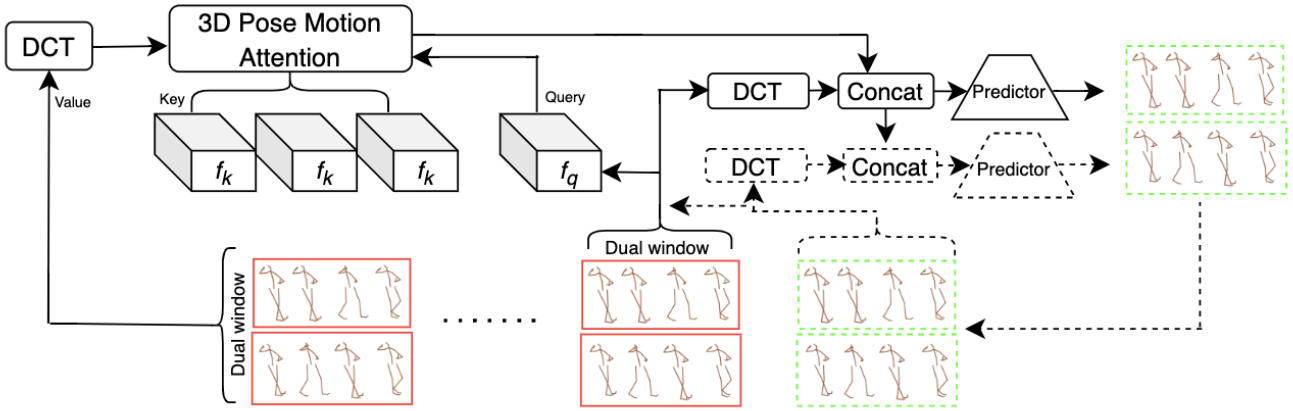

Figure 3: ***Overview of DeFT-Net****. Our re-timed DCT input poses are shown within the solid red boxes with the dual window extended history, and the predicted poses are shown within dotted green boxes. The last observed poses are initially used as query. For every consecutive poses in the history (key), we compute an attention score to weigh the dual window DCT coefficients (values) of the corresponding sub-sequence. The weighted sum of such values is then concatenated with the DCT coefficients of the last observed sub-sequence to predict the future. This comprises the transformer model of* OurDualWindowDCT.

flexibility in adjusting the relative positions of these slices. Our technique captures long-term temporal dependencies by taking into account different periods within human motion data, thus enhancing our model's ability to forecast future poses with an improved performance. As shown in Figure 2, we synthesize 3D pose data by interpolating frames containing motion foot anchor information from natural walking sequences in the Human 3.6M dataset. We then apply spherical interpolation to handle pose rotations and linear interpolation for pose translations to ensure smooth periodic motions. Since future frame forecasting from past sequences is crucial, our technique draws parallels with approaches such as those utilizing Discrete Cosine Transform (DCT) to encode motion, suppress high frequencies, and smooth jittery motions as seen in prior work [MLSL19, MLS20]. As the focus of our work is to adapt the attention model to periodic motion cycles, we fold the tensors that hold pose information to learn smooth motion transitions to form a dual-window stack model for improved short and long-term motion forecasting.

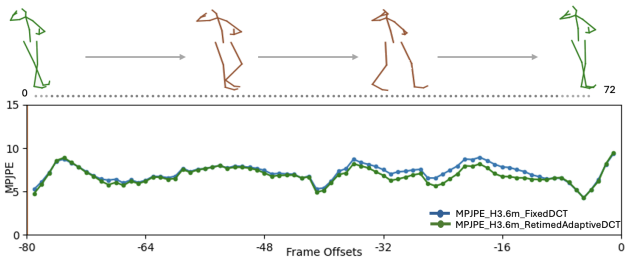

Figure 4: *From left-to-right, a plot visualisation of the Mean Per Joint Position Error (MPJPE) across 72 frames for training on both the History Repeats Itself DCT encoded motions and our dual-window extended DCT motion sequences*

3.1. Foot Anchor Frame Interpolation

As our goal is to learn from periodic walking sequence motions and forecast future pose motions, similar to *Cao et. al* [CGM*20], we rely on frame annotation based on the right foot placement at every n^{th} given frame. For periodic actions i.e walking and walking together, *linear interpolation* is applied to the root joint. As presented in equation 1, we compute a weighted average between the translation vectors of two key frames. Similar to [Kap15], we define a spherical path between the rotations and create key rotations from the rotation vectors of two consecutive frames as seen in equation 2 by *spherical linear interpolation*. We combine both interpolation techniques to achieve periodic dataset based foot anchor frame placements and pass these in an encoded DCT fashion to our dual window extended frequency transformer to learn from.

$$\text{lerp}(p_1, p_2, t) = (1-t)p_1 + tp_2 \tag{1}$$

$$\text{slerp}(q_1, q_2, t) = \frac{\sin((1-t)\theta)}{\sin(\theta)} q_1 + \frac{\sin(t\theta)}{\sin(\theta)} q_2 \tag{2}$$

3.2. Dual Window extended Frequency Attention Model

As our main aim is to take into account periodic pose sequences, we adapt a dual window by folding/stacking the tensors that hold the input features 3D pose sequence vectors along the bone axis. We then compute attention scores based on the *key* and *query* by feeding the encoded poses into a pytorch feed-forward function. Simi-

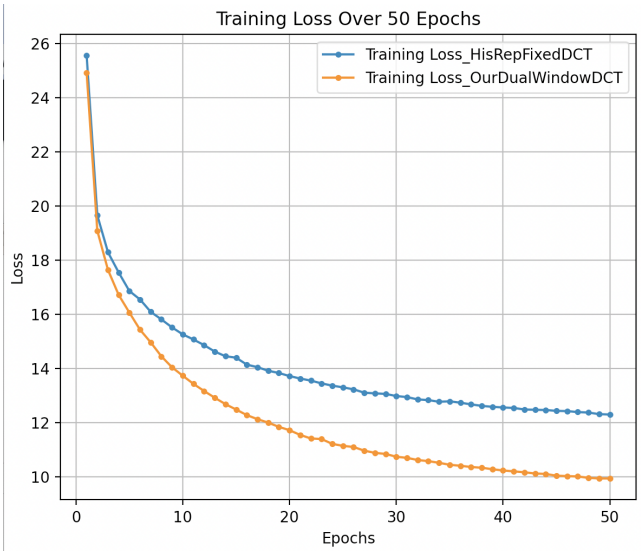

Figure 5: *Training Loss Comparison plot for 50 epochs: Dual window extended DCT vs History Repeats Itself fixed window DCT method [MLS20].*

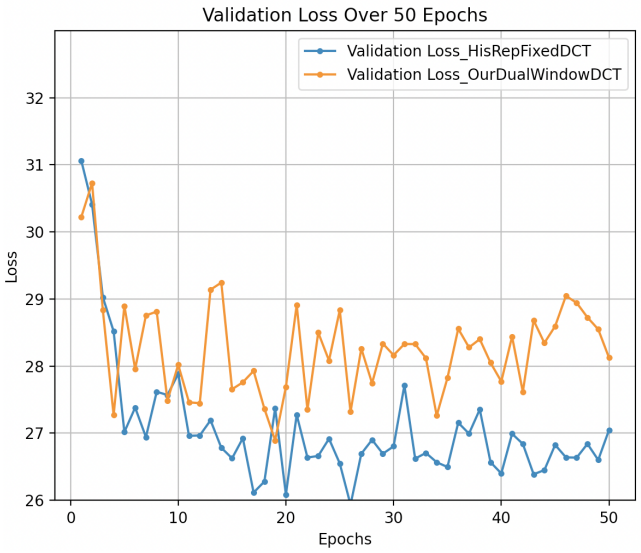

Figure 6: *Validation Loss Comparison plot for 50 epochs: Dual window extended DCT vs History Repeats Itself fixed window DCT method [MLS20].*

lar to *Mao et. al* [MLS20], we exploit motion attention as weights to aggregate our dual window extended DCT-encoded motion history into an estimate of future pose motion. This estimate is then combined with the latest observed motion, and the result then acts as input to a graph convolutional network (GCN), which lets our model better encode spatial dependencies between different joints. Our motion attention-based approach consistently outperforms the state-of-the art on short-term and long-term motion prediction by training a single unified model for both settings.

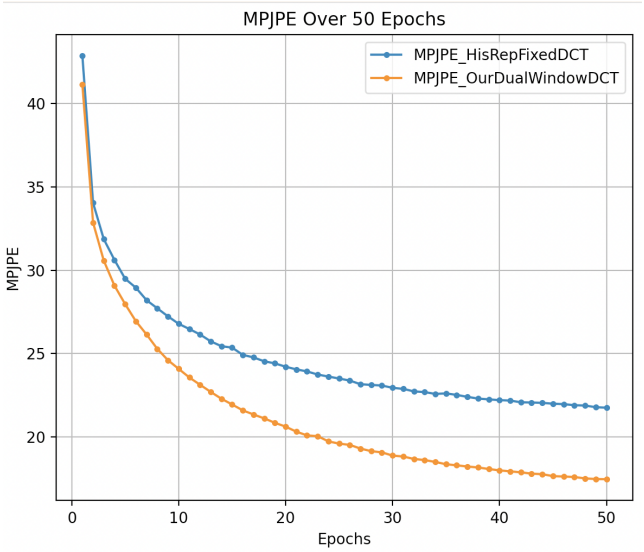

Figure 7: *MPJPE Comparison plot for 50 epochs: Dual window extended DCT vs History Repeats Itself fixed window DCT method [MLS20]*

	Walking			Walking Together		
Frame No.	1	3	5	8	9	10
HistRep [MLS20]	5.68	17.28	27.62	40.31	43.69	46.81
Ours	**5.45**	**16.78**	**26.50**	**38.41**	**41.69**	**44.78**

Table 1: *Following baseline setting MPJPE Batch evaluation results for test Subject 5 comparison on our re-timed interpolated vs original History Repeats Itself DCT [MLS20] method with Human3.6m datasets for predicting human motion at various frames for activities walking and walking together*

4. Experimental Results and Discussion

Following baselines setting [TMLZ18] [LZLL18], we present results for short-term and long-term predictions. On the H3.6M dataset, our dual-window DCT model is trained using a history of 50 frames to forecast the future 10 frames. Noticeably in Fig. 4, lower errors are observed at foot anchor frames where the right foot placements occur, and higher errors are noted when the feet are together during the mid-point of the cycle.

The results from the comparison of *HistRepeatDCT* and *OurDualWindowDCT*, highlight a trade-off between generalization and accuracy in joint position prediction. As shown in Fig 6, HistRepeatDCT exhibits a more stable and lower validation loss, suggesting it generalizes better to unseen data but is still comparable to our OurDualWindowDCT. This stability is crucial in ensuring that the model performs consistently well across different datasets and does not overfit to the training data. However, the higher MPJPE presented in Fig. 7 indicates that HistRepeatDCT is less accurate in predicting 3D joint pose positions, which can be a significant

drawback for tasks requiring precise motion forecasting, such as rhythmic motion prediction.

OurDualWindowDCT demonstrates superior performance in terms of training loss in Fig. 5 and MPJPE in Fig. 7, indicating better performance when capturing joint poses. This is essential for patterned human motion prediction, where precise joint movements are critical. However, the higher and more fluctuating validation loss in Fig. 6 points to potential overfitting and less reliable performance on new data which may also indicate limitations of the range of data used in training. Overall, despite its validation challenges, OurDualWindowDCT's lower MPJPE in our test results suggests it is a more robust model for rhythmic human motion prediction.

5. Conclusions and Future Work

In the paper, we have introduced a dual-windowed based motion attention model that exploits historical pose information according to the similarity between the current pose motion context and the cyclic sub-sequences in the pose motion history. Our approach achieves state-of-the-art performance in predicting rhythmic motion by re-timing the Human3.6m dataset based on foot anchor placements. Furthermore, our experiments have demonstrated that our network generalizes to previously unseen *walking* and *walking-together* motion sequences. To leverage the strengths of *OurDualWindowDCT* while mitigating its drawbacks, further techniques such as cross-validation, regularization, or data augmentation would be employed to enhance its generalization ability while maintaining its predictive accuracy. We aim to further investigate the use of a stack motion attention model to discover human motion patterns in body parts, such as legs, to get more flexible attention for a DCT history of multiple pose periods and consideration of non-linear re-timing approaches.

Acknowledgements

This work is supported by funding EU's Horizon 2020 research and innovation programme under grant agreement No. 101017779.

References

[AKCH21] AKSAN E., KAUFMANN M., CAO P., HILLIGES O.: A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)* (2021), IEEE, pp. 565–574. 3

[AKH19] AKSAN E., KAUFMANN M., HILLIGES O.: Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 7144–7153. 2

[ANBH23] ALEXANDERSON S., NAGY R., BESKOW J., HENTER G. E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph. 42*, 4 (2023), 44:1–44:20. doi:10.1145/3592458. 3

[BBKK17] BUTEPAGE J., BLACK M. J., KRAGIC D., KJELLSTROM H.: Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 6158–6166. 2

[BH00] BRAND M., HERTZMANN A.: Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (USA, July 2000), SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., pp. 183–192. URL: https://dl.acm.org/doi/10.1145/344779.344865, doi:10.1145/344779.344865. 2

[CAW*19] CHIU H.-K., ADELI E., WANG B., HUANG D.-A., NIEBLES J. C.: Action-agnostic human pose forecasting. In *2019 IEEE winter conference on applications of computer vision (WACV)* (2019), IEEE, pp. 1423–1432. 2

[CGM*20] CAO Z., GAO H., MANGALAM K., CAI Q.-Z., VO M., MALIK J.: Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16* (2020), Springer, pp. 387–404. 4

[CHW*20] CAI Y., HUANG L., WANG Y., CHAM T.-J., CAI J., YUAN J., LIU J., YANG X., ZHU Y., SHEN X., ET AL.: Learning progressive joint propagation for human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16* (2020), Springer, pp. 226–242. 3

[CSY20] CUI Q., SUN H., YANG F.: Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 6519–6527. 2

[FLFM15] FRAGKIADAKI K., LEVINE S., FELSEN P., MALIK J.: Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 4346–4354. 2

[FYD*23] FU J., YANG F., DANG Y., LIU X., YIN J.: Learning constrained dynamic correlations in spatiotemporal graphs for motion prediction. *IEEE Transactions on Neural Networks and Learning Systems* (2023). 2

[GMI23] GUINOT L., MATSUMOTO R., IWATA H.: Stacked dual attention for joint dependency awareness in pose reconstruction and motion prediction. 3

[GSAH17] GHOSH P., SONG J., AKSAN E., HILLIGES O.: Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)* (2017), IEEE, pp. 458–466. 2

[GWLM18] GUI L.-Y., WANG Y.-X., LIANG X., MOURA J. M.: Adversarial geometry-aware human motion prediction. In *Proceedings of the european conference on computer vision (ECCV)* (2018), pp. 786–803. 2

[HGMN19] HERNANDEZ A., GALL J., MORENO-NOGUER F.: Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 7134–7143. 2

[IPOS13] IONESCU C., PAPAVA D., OLARU V., SMINCHISESCU C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence 36*, 7 (2013), 1325–1339. 2, 3

[JZSS16] JAIN A., ZAMIR A. R., SAVARESE S., SAXENA A.: Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition* (2016), pp. 5308–5317. 2, 3

[Kap15] KAPOULKINE A.: *Approximating slerp*, 2015. URL: https://zeux.io/2015/07/23/approximating-slerp/. 4

[KSM24] KONIARIS B., SINCLAIR D., MITCHELL K.: Dancemark: An open telemetry framework for latency-sensitive real-time networked immersive experiences. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2024), IEEE, pp. 462–463. 2

[LMB14] LOPER M., MAHMOOD N., BLACK M. J.: Mosh: motion and shape capture from sparse markers. *ACM Trans. Graph. 33*, 6 (2014), 220–1. 3

[LZLL18] LI C., ZHANG Z., LEE W. S., LEE G. H.: Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 5226–5234. 2, 3, 5

[MBR17] MARTINEZ J., BLACK M. J., ROMERO J.: On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 2891–2900. 2

[MGT*19] MAHMOOD N., GHORBANI N., TROJE N. F., PONS-MOLL G., BLACK M. J.: Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 5442–5451. 3

[MLS20] MAO W., LIU M., SALZMANN M.: History repeats itself: Human motion prediction via motion attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16* (2020), Springer, pp. 474–489. 2, 3, 4, 5

[MLSL19] MAO W., LIU M., SALZMANN M., LI H.: Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 9489–9497. 2, 3, 4

[PGA18] PAVLLO D., GRANGIER D., AULI M.: Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485* (2018). 2

[SAKM23] SINCLAIR D., ADEMOLA A. V., KONIARIS B., MITCHELL K.: Dancegraph: A complementary architecture for synchronous dancing online. In *36th International Computer Animation Social Agents (CASA) 2023* (2023). 2

[THR06] TAYLOR G. W., HINTON G. E., ROWEIS S.: Modeling Human Motion Using Binary Latent Variables. In *Advances in Neural Information Processing Systems* (2006), vol. 19, MIT Press. URL: https://proceedings.neurips.cc/paper/2006/hash/1091660f3dff84fd648efe31391c5524-Abstract.html. 2

[TMLZ18] TANG Y., MA L., LIU W., ZHENG W.: Long-term human motion prediction by modeling motion context and enhancing motion dynamic. *arXiv preprint arXiv:1805.02513* (2018). 5

[VPHB*21] VALLE-PÉREZ G., HENTER G. E., BESKOW J., HOLZAPFEL A., OUDEYER P.-Y., ALEXANDERSON S.: Transflower: Probabilistic autoregressive dance generation with multimodal attention. *ACM Trans. Graph. 40*, 6 (2021), 195:1–195:14. doi:10.1145/3478513.3480570. 3