AI

MDPI

Article

Dynamic Differencing-Based Hybrid Network for Improved 3D Skeleton-Based Motion Prediction

Ruiya Ji [1], Chengjie Lu [2] and Jianqi Zhong [2,*]

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK; r.ji@se24.qmul.ac.uk
[2] College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China; 2022280594@email.szu.edu.cn
* Correspondence: zhongjianqi2017@email.szu.edu.cn

Abstract: Background: Three-dimensional skeleton-based human motion prediction is an essential and challenging task for human–machine interactions, aiming to forecast future poses given a history of previous motions. However, existing methods often fail to effectively model dynamic changes and optimize spatial–temporal features. **Methods:** In this paper, we introduce Dynamic Differencing-based Hybrid Networks (2DHnet), which addresses these issues with two innovations: the Dynamic Differential Dependencies Extractor (2D-DE) for capturing dynamic features like velocity and acceleration, and the Attention-based Spatial–Temporal Dependencies Extractor (AST-DE) for enhancing spatial–temporal correlations. The 2DHnet combines these into a dual-branch network, offering a comprehensive motion representation. **Results:** Experiments on the Human3.6M and 3DPW datasets show that 2DHnet significantly outperforms existing methods, with average improvements of 4.7% and 26.6% in MPJPE, respectively.

Keywords: human motion prediction; dynamic difference; graph convolutional network; multi-layer perceptron

check for updates

Citation: Ji, R.; Lu, C.; Zhong, J. Dynamic Differencing-Based Hybrid Network for Improved 3D Skeleton-Based Motion Prediction. *AI* **2024**, *5*, 2897–2913. https://doi.org/10.3390/ai5040139

Academic Editor: Giovanni Diraco

Received: 11 September 2024
Revised: 5 November 2024
Accepted: 3 December 2024
Published: 11 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Human motion prediction is a key task in human–computer interaction and computer vision, with vital applications in areas such as security surveillance [1], human–robot interaction [2], and virtual gaming [3]. The aim is to predict future human poses from past movements with accuracy. This task is challenging because of the high dimensionality, complexity, and dynamic nature of human motion.

In the past, human motion prediction models have been built using Gaussian process latent variable models [4] and hidden Markov models [5]. However, there has been a recent shift towards deep learning approaches, leveraging a variety of neural networks, including recurrent neural networks (RNNs) [6–9], graph convolutional networks (GCNs) [10–15], Transformers [16], and multi-layer perceptrons (MLPs) [17–19].

Although promising results have been achieved in previous works, existing approaches do not efficiently address two main challenging factors. Firstly, existing methods often struggle to effectively model rapid changes or abrupt dynamics in actions, ignoring the effectiveness of the dynamic characteristics for motion prediction; e.g., they capture the acceleration when jumping from a standing position or deceleration when coming to a sudden stop. Acceleration information helps the model to understand abrupt changes or dynamic patterns in motion. Secondly, previous methods typically focus on optimizing a single type of feature or refining specific network architectures to enhance the extraction of spatial or temporal features.

To address the first issue, we propose the Dynamic Differential Dependencies Extractor (2D-DE), which is inspired by the observation that differential operation can capture

dynamic features for human motion modeling. For example, first-order temporal differencing calculates the positional changes in skeletal key points between consecutive frames, effectively capturing the velocity of each key point. Velocity provides insights into the speed and dynamic variations in the motion, such as the swinging speed of an arm or the leg movement speed while walking, which is valuable for understanding and predicting the progression of the action. Additionally, second-order temporal differencing further computes the acceleration of skeletal key points, which reveals the acceleration or deceleration patterns in the motion. For example, it captures the acceleration when jumping from a standing position or deceleration when coming to a sudden stop. Acceleration information helps the model to understand abrupt changes or dynamic patterns in motion.

For the second issue, we develop an Attention-based Spatial–Temporal Dependencies Extractor (AST-DE) to capture compact and effective motion dependencies at the spatial and temporal levels. In AST-DE, we propose Temporal-wise Attention and Spatial-wise Feature Extractor to fully exploit the spatial–temporal kinematic correlations, respectively.

Building on these two modules, we propose Dynamic Differencing-based Hybrid Networks (2DHnet), a novel model that uses a dual-branch network to learn motion representations for improved prediction (see Figure 1). 2DHnet aims to combine spatiotemporal dependency features with dynamic differential features, which provides a more holistic representation of human motion, encompassing both static skeletal structure and dynamic motion changes. Extensive experiments on two large-scale datasets, Human3.6M and 3DPW, demonstrate that our model outperforms most state-of-the-art methods in both short-term and long-term predictions regarding effectiveness and efficiency. The main contributions of this paper are as follows:

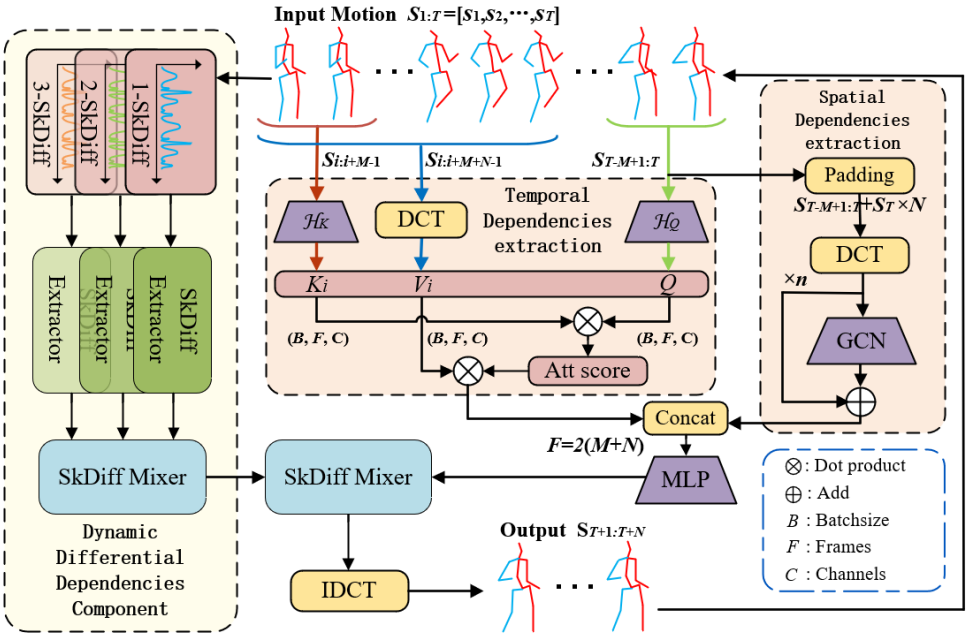

Figure 1. Network architecture of our proposed method. The entire network structure is divided into two components: Dependencies Extractor (AST-DE) in pink blocks and Dynamic Differential Dependencies Extractor (2D-DE) in yellow block. Specifically, AST-DE is the main branches including Temporal-wise Attention and Spatial-wise Feature Extractor, while 2D-DE is a differential branch, which is used to extract differential information from sequences and merge it with features extracted from other branches. Temporal-wise Attention extracts temporal features of the sequence, dividing the input historical sequence into three parts. The green parentheses represent the *query*, while the red and blue parentheses represent the *key* and *value*, respectively. Spatial-wise Feature Extractor is used to extract features about the human body in spatial dimensions, mainly using graph convolutional networks.

- We propose Dynamic Differencing-based Hybrid Networks (2DHnet), providing a more holistic representation of human motion, encompassing both static skeletal structure and dynamic motion changes, and enabling effective human motion prediction.
- In 2DHnet, we develop the Attention-based Spatial–Temporal Dependencies Extractor (AST-DE), which includes two main components: (1) Temporal-wise Attention, to capture features from historical information in the time dimension; and (2) Spatial-wise Feature Extractor, to capture spatial features between human joints.
- In 2DHnet, we develop the Dynamic Differential Dependencies Extractor (2D-DE), which leverages multiple differential operations on 3D skeleton data to extract rich dynamic features, enabling 2DHnet to more accurately capture dynamic information such as velocity and acceleration, which enhances prediction accuracy.
- We conduct experiments to quantitatively and qualitatively verify that our proposed 2DHnet consistently outperforms existing methods, by 4.7%, 26.6% of MPJPE on average on the Human3.6M and 3DPW datasets, respectively.

The rest of this paper is organized as follows: We review the related work on 3D human motion prediction most relevant to our approach in Section 2. Section 3 mainly introduces the overall framework and method. Section 4 describes the details of the experimental implementation and test results on the two datasets to demonstrate the effectiveness of our model. Section 5 concludes the work.

2. Related Work

2.1. Three-Dimensional Skeleton-Based Human Motion Prediction

The extensive research into neural networks has catalyzed significant progress in the domain of motion forecasting based on 3D skeletal data. Among various predictive models, recurrent neural networks (RNNs) stand out as a prevalent choice for tackling tasks involving the prediction of sequences, such as anticipating the trajectory of human body movements. These networks are particularly adept at discerning patterns within temporal data, which is crucial for accurately simulating the future positions of skeletal joints. Li et al. [20] introduced a human motion prediction approach grounded in a convolutional sequence-to-sequence model, leveraging RNNs to capture the temporal characteristics of human movement sequences. Li et al. [21] proposed an independent recurrent neural network architecture designed to construct longer and deeper RNN models, thereby enhancing the performance of human motion prediction. Liu et al. [22] presented an LSTM network based on global context awareness from skeletal data. Tang et al. [23] combined RNNs with an attention mechanism for human pose prediction. This approach takes into account the spatial relationships between different joints and their temporal correlations, utilizing the query, key, and value of the attention mechanism to select the necessary information for subsequent predictions.

RNNs have traditionally shown their strength in predicting the temporal aspects of motion, leveraging their ability to process sequences and recognize patterns over time. In contrast, graph convolutional networks (GCNs) have been particularly adept at analyzing spatial relationships within data, effectively handling the structural aspects of motion by operating on graphs that represent the spatial configuration of entities. Zhong et al. [24] introduced a spatio-temporal gated adjacency GCN, which captured complex spatio-temporal dependencies by balancing the weighting of spatial and temporal modeling and integrating decoupled spatio-temporal features. He et al. [25] proposed a two-stage model for human motion prediction that builds on enhanced GCNs. This model first produces preliminary predictions through spatial attention graph convolutional layers, then refines these predictions with causal temporal graph convolutional layers to improve accuracy. Fu et al. [26] developed a model that applies GCNs on spatiotemporal graphs to capture dynamic spatiotemporal dependencies among human joints, thereby predicting motion trajectories. To achieve a more detailed and enriched representation of joint connections, Gu et al. [14] introduced a dual-path GCN framework that learns the GCN adjacency matrices of two paths interactively, capturing the correlations between joint positions and velocities.

Transform-based models like Transformers offer parallel processing and enhanced spatio-temporal feature capture. Martínez-González et al. [27] proposed Pose Transformer (POTR), a non-autoregressive Transformer architecture capable of decoding elements in the query sequence in parallel, thereby reducing computational complexity and potentially preventing error propagation in long-term predictions. Mi et al. [28] presented a Transformer network that integrates spatial and positional encoding for human motion prediction from skeletal data. This approach enhances GCNs through the incorporation of a deep, multilayered residual graph framework, effectively reducing over-smoothing effects and more precisely modeling spatial relationships. Zhao et al. [29] introduced a Bidirectional Transformer GAN (BiTGAN), utilizing bidirectional Transformer encoders and decoders to capture historical motion information and employing Soft-DTW loss to maintain similarity between predictions and actual motions.

2.2. MLP-Based Feature Modeling

The employment of MLP-based methods for feature extraction has demonstrated their prowess in unearthing intricate data patterns through the intricate interplay of neurons within their layered frameworks.

MLP has been widely applied in the field of object detection. Menese et al. [30] proposed SmartSORT for real-time multi-object tracking. The method leverages the powerful feature extraction capabilities of MLPs to achieve fast and accurate tracking of multiple targets. Cao et al. [31] employed a Query-Independent Category Supervision (QICS) method for modeling category information and introduced a deep MLP to capture both long-range and short-range information, which integrated a deep MLP into a detection framework based on Transformers. Chen et al. [32] proposed CycleMLP based on MLP for dense visual prediction tasks such as object detection, and enhanced the transmission and fusion of features through recurrent connections, thereby improving detection performance.

MLP has also been applied to facial recognition. Boughrara et al. [33] proposed a constructive training algorithm, which incrementally adds hidden neurons to the network in a manner that enhances its ability to learn and generalize from facial data. Shahreza et al. [34] aimed to address privacy protection in facial recognition and proposed MLP-Hash, a method that transforms facial templates into irreversible and unlinkable forms by passing them through a randomized MLP and then hashing the output. This approach ensures that sensitive biometric data remain secure while still allowing for recognition tasks.

In addition, MLP has been proven effective in the field of 3D skeleton-based motion analysis. Guo et al. [19] proposed a network only using MLP to extract the spatial and temporal features, respectively, and proved that MLP can efficiently extract patterns from human historical pose sequences, greatly reducing the complexity and parameter count of the model. As a classic approach to feature extraction, the sustained prominence of MLPs in the field of computer vision indicates a continuous evolution, characterized by ongoing enhancements that tailor MLPs to meet the shifting complexities of advanced predictive analytics. Bouazizi et al. [35] introduced Mixer, using a spatial MLP to extract fine-grained spatial dependencies of body joints and a temporal MLP to model their interactions over time, ultimately aggregating these spatial–temporal mixed features to predict future human poses.

3. Method

3.1. Overall Architecture

Our task is to forecast the future sequence of human poses based on the given past poses, using multiple frames of 3D coordinates to represent an action. Specifically, let $s_t \in \mathbb{R}^{3\times K}$ represent the pose matrix at the t-th frame, where K denotes the number of body joints. $S_{1:T} = [s_1, s_2, ..., s_T]$ denotes T consecutive frames of historical human poses and the goal is to predict the future N frames of the actions, which are denoted as $S_{T+1:T+N} = [s_{T+1}, s_{T+2}, ..., s_{T+N}]$.

Figure 1 indicates the whole structure of our network, which consists of two key components: Attention-based Spatial–Temporal Dependencies Extractor (AST-DE) to capture compact and effective motion dependencies at spatial and temporal levels, and the Dynamic Differential Dependencies Extractor (2D-DE) to capture the dynamic characteristics of movements. In AST-DE, the model learns the temporal and spatial features of action sequences. Inspired by Mao et al. [13], we propose Temporal-wise Attention for comparing similarities in historical sequences and extracting temporal information. In the Spatial-wise Feature Extractor, we use a GCN to extract the position information and correlation degree of joint points in human motion. For 2D-DE, we devised a technique to compute the differences within a given sequence. We prepend one frame to the beginning of the resulting sequence to maintain consistent data dimensions. This process is repeated P times, yielding P distinct series that capture various dynamics of human motion, such as velocity, acceleration, and finer motion variations. Finally, we merge the information extracted from the two branches to obtain a more complete and representative feature representation. Throughout the process, we utilize Discrete Cosine Transform (DCT) [12] and Inverse Discrete Cosine Transform (IDCT) transformations, transforming the input sequence from the time domain to the frequency domain, representing the original sequence with DCT coefficients to improve the robustness and generalization ability of the model, and finally using IDCT to transform the sequence back to the time domain as the output.

3.2. Attention-Based Spatial–Temporal Dependencies Extractor (AST-DE)

This section of the network separately extracts spatial and temporal features from the sequence in Euclidean space and then fuses them. The model takes an input sequence of T frames representing an action and outputs a predicted sequence of N frames. It comprises two main components: (1) Temporal-wise Attention, which takes advantage of Attention mechanism and extracts features from historical information along the time dimension; and (2) Spatial-wise Feature Extractor, which captures positional relationships such as connectivity and other spatial characteristics between human joints.

3.2.1. Temporal-Wise Attention

To recognize similarities and repetitions in human motion sequences, we apply a Temporal-wise Attention. This technique determines the attention score by matching the current frame with historical motion frames through a *query* and various *key–value* pairs, allowing us to identify similar frames. The input historical sequence is divided into three parts: *query*, *key*, and *value*. For a sequence, there is only one *query*, the *key* and *value* slide on the sequence, allowing us to count the correlation for each segment. The *query* is the last M frames of the input, used to calculate the similarity with the sequence before it. In order to determine attention weights with the *query*, the *key* is also set to M consecutive frames, and it has $T - M - N + 1$ in the entire sequence. The *query* and *key* are, respectively, represented as $S_{T-M+1:T}$ and $\{S_{i:i+M-1}\}_{i=1}^{T-N-M+1}$, mapped to the unified dimension using their respective mapping functions $\mathcal{H}_Q(\cdot)$ and $\mathcal{H}_K(\cdot)$, that is,

$$Q = \mathcal{H}_Q(S_{T-M+1:T}) \tag{1}$$

$$K_i = \mathcal{H}_K\left(\{S_{i:i+M+N-1}\}_{i=1}^{T-N-M+1}\right) \tag{2}$$

The attention score between each frame of these two sequences is formulated as

$$a_i = \frac{Q \times K_i}{\sum_{i=1}^{T-N-M+1} Q \times K_i} \tag{3}$$

where a_i denotes the attention score. We utilize sum normalization and the ReLU [36] activation function to constrain the weight values within the range of 0 and 1.

For each *key*, there is a corresponding *value*. In this work, *value* contains the *key* and the subsequent N frames following it, with the length of $N + M$, denoted as

$\{S_{i:i+M+N-1}\}_{i=1}^{T-N-M+1}$. The same as for the *key*, there is also $T-M-N+1$ *values* in the entire sequence. Before weighting, we first apply DCT to transform it from the time domain to the frequency domain, and use the frequency-domain information V_i to encode the features of the sequence. Then, the model yields the sequence S^{tem} by performing the dot product between V_i and a_i, that is,

$$S^{tem} = \sum_{i=1}^{T-N-M+1} a_i V_i \tag{4}$$

3.2.2. Spatial-Wise Feature Extractor

GCN demonstrates exceptional adaptability to various poses and deformations in human motion due to its flexibility in handing graph-structured data. It effectively captures spatial relationships and connectivity patterns between human joints, which are essential for understanding the structure and semantics of human motion. Therefore, we utilize stacked GCN blocks to extract features only in the spatial dimension. Before this, we pad the *query* sequence $S_{T-M+1:T}$ with N frames of the last frame and convert it to the frequency domain using DCT. Mathematically, $A^{(j)} \in \mathbb{R}^{K\times K}$ represents the trainable graph adjacency matrix, where j denotes the number of the layer. At the j-th layer, the graph convolution is

$$G^{(j+1)} = \sigma(A^{(j)} \times G^{(j)} \times W^{(j)}) \tag{5}$$

where $\sigma(\cdot)$ denotes the activation function $tanh(\cdot)$, and $W^{(j)}$ denotes the trainable weights of the j-th layer, which effectively enforces limitations on feature values to maintain consistent predictions.

We subsequently concatenate the outputs S^{tem} and S^{spa}, then utilize fully connected layers to blend the spatial and temporal features, with the goal of enhancing the representation of human joint coordinates. After IDCT, the sequence returns to the time-domain space, and we take the last $N+M$ frames $\hat{S}$ as the output of the network.

3.3. Principle of SkDiff

Paralleling with the AST-DE method, the Dynamic Differential Dependencies Extractor empowers the effectiveness of the prediction. As for the building method SkDiff, we take frames with a length of T as our input. As the series is discontinuous, the N-th initial difference block D_0^N can simply be given by

$$D_0^N = X_{2:T}^N - X_{1:T-1}^N \tag{6}$$

However, the length of the generated series is $T-1$, which does not correspond to the required input series length of our model. Thus, we developed two methods for difference series compensation: (a) Repeat the first frame at the beginning of the series, and (b) repeat the last frame at the end of the series. The former method focuses on the vitality of the recent motion, while the latter concentrates on stimulating the model capacity of eliminating inaccurate prediction. Based on the two methods, we can refine the series for difference feature extraction as D^N:

$$D^N = \begin{cases} [x_1, x_1, x_2, x_3, ..., x_{T-2}, x_{T-1}] & \text{(a)} \\ [x_1, x_2, x_3, ..., x_{T-2}, x_{T-1}, x_{T-1}] & \text{(b)} \end{cases} \tag{7}$$

where x_i represents the i-th frame in the original difference block D_0^N. Therefore, we can ensure that the length of the input D^N is T, which means our difference extractor can grasp the feature from the difference block. Specifically, our difference extractor is named as SkDiff Extractor, whose structure is demonstrated in Figure 2. For corresponding to the data feature, our SkDiff Extractor can be flexibly alternated from the given models.

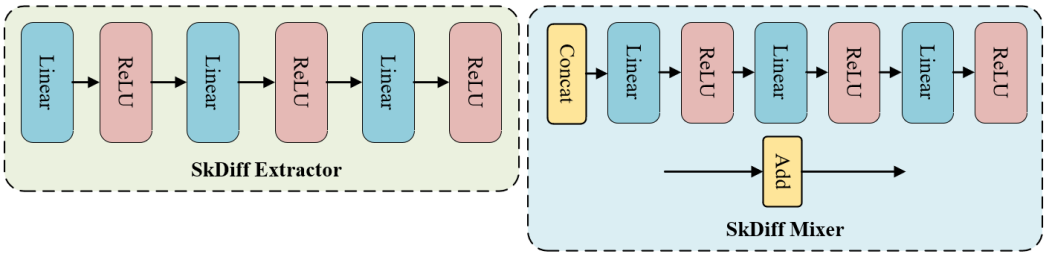

Figure 2. The structure of SkDiff Extractor and Mixer in 2D-DE. SkDiff Extractors are based on multi-layer perceptron and consist of linear layers and an activation function (ReLU). Mixer is dynamic, corresponding to the data feature, which is chosen from concatenating the input data together before fusing features or adding all results of SkDiff Extractors.

SkDiff can generate blocks of various difference depths, which facilitates the advantage that it can handle a range of circumstances varying in the extent of motion change. In Figure 1, we demonstrate that the 1-SkDiff block is generated by SkDiff performed on the original motion series, while the 2-SkDiff block is yielded by SkDiff performed on the 1-SkDiff block. The depths of SkDiff generation can be adjusted according to the motion characteristics in reality.

3.4. Loss

The model yields a predict sequence with a total length of $M + N$. For accuracy assessment, we measure the deviation between the 3D spatial coordinates of each human body joint as predicted by the model and the actual coordinates. To achieve this, we adopt the mean per-joint position error (MPJPE), which was initially suggested in [12] for the estimation of 3D joint positions. This metric quantifies the mean squared Euclidean distance for each joint's predicted location in comparison with the ground truth location, and is expressed by the following equation:

$$\mathcal{L} = \frac{1}{K} \sum_{k=1}^{K} ||\hat{J}_k - J_k||^2 \tag{8}$$

where $\hat{J}_k$ represents the 3D coordinates of the k-th joint of the human pose, and J_k denotes the ground truth.

4. Experimental Results and Datasets

4.1. Datasets

In order to evaluate the effectiveness of the proposed 2DHnet in the human pose estimation task, we utilize publicly available datasets as benchmarks, specifically Human3.6M (H3.6M) [37] and 3D Pose in the Wild (3DPW) [38].

4.1.1. Human3.6M

Human3.6M [37] stands as a pivotal and extensively utilized resource in the field of 3D human pose estimation and human motion prediction. Encompassing over 3.6 million images, this dataset was meticulously captured using the advanced Vicon motion capture system. It features seven actors performing a diverse array of 15 distinct activities, ranging from routine actions such as walking and smoking to more complex interactions like greeting and discussing.

Each pose within the dataset is anatomically defined by 32 joints, which are meticulously pre-processed into 3D coordinates. To facilitate a more consistent temporal resolution, the sequences have been uniformly down-sampled to 25 frames per second. This down-sampling aids in streamlining the data for computational processes while preserving the essential dynamics of human motion.

Following established research routines, we make use of subjects 1 (S1), S6, and S7–S9 for training, S11 for validation, and S5 for testing. This methodology aligns with the conventions observed in previous studies, thereby promoting standardization in evaluation metrics across the research community.

4.1.2. 3DPW

The 3D Pose in the Wild (3DPW) dataset [38] stands as a significant resource in the field of 3D human pose estimation, particularly known for its comprehensive capture of natural human movements in diverse and complex environments. It is one of the first datasets to offer precise 3D poses in outdoor scenarios, making it an essential benchmark for evaluating models in realistic settings. The dataset includes over 51,000 frames across 60 video sequences. Each human pose in 3DPW is articulated through 24 to 26 joints, with variations depending on specific evaluation protocols. The dataset has been divided into training, validation, and test sets. The high frame rate of 60 Hz further adds to its suitability for dynamic and real-time pose estimation tasks. Following a common practice adopted by most studies, we train our models over 50 epochs on the training set before testing on the designated test set.

4.2. Experimental Metric

To evaluate the performance of our model, which outputs 3D positions, we adopted the mean per-joint position error (MPJPE) metric on the 3D joints. The MPJPE is the most widely used metric for assessing errors in 3D pose estimation, as it measures the discrepancies in 3D coordinate space. For an integrated comparison with other methods, we adhered to the evaluation standard outlined in [12,13,39,40], applying a sampling rate of 25 frames per second across both datasets. The L2 distance between the predicted and ground truth positions of each joint was calculated and then averaged to quantify the error.

4.3. Experimental Settings

In order to implement the experiment, we made a series of parameter settings to enhance the model effectiveness. Accordingly, we performed experiments and obtained the most effective parameter configuration for our model.

Within the processing of both datasets, the input sequence length was established at 50 frames. For the Human3.6 dataset, the output was configured to 10 frames, whereas for the 3DPW dataset, it was set to 25 frames, for this dataset is more changeable and needs more future ground truths to generate more transitive future predictions. Regarding the feature representation of the human body, each joint is described by three-dimensional coordinates, represented as $C = 3 \times K$. Specifically, the Human3.6 dataset employs 22 joints, resulting in a feature dimension of 66. We kept the same number of joints for 3DPW.

As is elaborated in Section 3.1, for the 2D-DE, we built a method which can yield the difference of the given series, while we padded one frame at the beginning of the generated series to ensure the same data dimensions. We performed this method P times and thus obtained P series, which in reality meant velocity, acceleration, and more slight changes of human motion. Then, we transformed the temporary skeleton series to frequency field data utilizing DCT. For feature extraction of each difference layer, we adopted the network of the multi-layer perceptron (MLP) block, where the latent layer size was 128. Then, these layers were concatenated and merged with another MLP block, where the latent layer size was $128 \times P$. Finally, the difference feature output was concatenated with the attention-based feature output, mixed with a 256-latent-sized MLP block, and transformed into temporary skeleton-based series as the ultimate prediction.

For feature extraction from each difference layer, we utilized a multi-layer perceptron (MLP) network, where the latent layer size was set to 128. The outputs of these layers were then concatenated and merged using another MLP block, with a latent layer size of $128 \times P$. The final difference features were combined with the attention-based features, passed through an MLP block with a latent size of 256, and ultimately converted back into a temporal skeleton-based sequence utilizing IDCT, providing the final prediction.

According to Section 3.2, the Temporal-wise Attention learning process is integrated with the attention mechanism-based motion feature abstracting method. Practically, the *query* and *key* lengths were fixed at 10 frames, denoted as $M = 10$. The mappings $\mathcal{H}_Q(\cdot)$ and $\mathcal{H}_K(\cdot)$ project Q and K_i into 256 dimensions for attention weight computation. The *value* length was configured to 20, meaning that the model predicted the most pertinent 20 frames based on the last 10 frames of the input sequence. During the padding phase, the final frame of the input sequence was replicated 10 times to generate a total of 20 frames. These frames were then concatenated with the preceding 20 frames, forming a 40-frame input for the GCNs.

In the learning phase of the Spatial-wise Feature Extractor, the dimensions of W were established as 256×256, while A had dimensions of $C \times C$. We applied 18 GCN blocks in a stacked arrangement, connected via residual structures to improve generalization performance and accelerate model concentration.

To train the network, the Adam optimizer was employed with a batch size of 32. The model, comprising 15.34 million parameters, was implemented using the PyTorch framework and trained over 50 epochs on one NVIDIA RTX 4070 GPU, with the initial learning rate set to 0.001.

4.4. Comparison with State-of-the-Art Methods

In this section, we systematically evaluate the performance of various human pose estimation methods, with a particular focus on comparing them to our newly proposed model. To ensure a comprehensive analysis, we report the results of our model across both short-term (0∼500 ms) and long-term (500∼1000 ms) prediction intervals, as these time frames provide insights into the model's performance across different temporal horizons.

For the Human3.6M dataset, our model is trained using the past 50 frames to predict the subsequent 10 frames. Once the next 10 frames are predicted, these frames are iteratively fed back into the model as input to generate further predictions. This recursive approach allows for the evaluation of our model's ability to maintain accuracy over multiple steps of prediction. We note that the SkDiff depth represents the number of iterations of SkDiff. Among the results of the experiments, we find that the trained model achieves the lowest MPJPE when the SkDiff depth is 1.

On the 3DPW dataset, a similar methodology is adopted, with the model utilizing the past 50 frames to predict the next 25 frames. This setting provides a broader time window for prediction, enabling a more thorough examination of the model's robustness over longer-term forecasts.

4.4.1. Results on Human3.6M

The performance of our model on the Human3.6M dataset, as presented in Tables 1 and 2, showcases its superiority in both short-term and long-term 3D motion prediction, measured by mean per-joint position error (MPJPE). Our model not only competes effectively with state-of-the-art methods like TFAN [17] and EqMotion [18] but also demonstrates distinct advantages in capturing both immediate and sustained motion trends.

Table 1. Short-term results on the 15 actions of H3.6M dataset. The average prediction results for all the actions are added at the end. The best results are highlighted in bold.

Motion	**Walking**				**Eating**				**Smoking**				**Discussion**			
Milliseconds	**80**	**160**	**320**	**400**	**80**	**160**	**320**	**400**	**80**	**160**	**320**	**400**	**80**	**160**	**320**	**400**
PGBIG [40]	10.2	19.8	34.5	40.3	7.0	15.1	30.6	38.1	6.6	14.1	28.2	34.7	10.0	23.8	53.6	66.7
SPGSN [41]	10.1	19.4	34.8	41.5	7.1	14.9	30.5	37.9	6.7	13.8	28.0	34.6	10.4	23.8	53.6	67.1
EqMotion [18]	**9.2**	**18.2**	34.2	41.4	6.6	14.2	30.0	37.7	6.2	12.8	26.7	33.6	9.1	22.0	51.8	65.3
SIMLPE [19]	9.9	-	-	39.6	5.9	-	-	36.1	6.5	-	-	36.3	9.4	-	-	64.3
TFAN [17]	10.0	19.9	36.5	43.1	5.9	13.6	28.4	35.9	6.6	14.3	29.8	36.8	9.1	21.9	50.0	63.4
April-GCN [14]	**9.2**	18.7	34.4	41.1	6.2	14.1	30.0	37.7	**5.8**	13.1	27.0	33.7	8.9	22.6	52.8	65.8
Ours	9.5	18.9	**33.0**	**38.1**	**5.7**	**13.3**	27.9	**35.2**	5.9	**12.3**	**23.9**	**30.1**	**8.2**	**17.7**	**34.1**	**43.4**

Table 1. *Cont.*

Motion	**Directions**				**Greeting**				**Phoning**				**Posing**			
Milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
PGBIG [40]	7.2	17.6	40.9	51.5	15.2	34.1	71.6	87.1	8.3	18.3	38.7	48.4	10.7	25.7	60.0	76.6
SPGSN [41]	7.4	17.2	39.8	50.3	14.6	32.6	70.6	86.4	8.7	18.3	38.7	48.5	10.7	25.3	59.9	76.5
EqMotion [18]	6.6	**15.8**	**39.0**	**50.0**	12.7	30.0	69.1	86.3	7.7	**17.3**	38.2	48.4	9.3	23.7	59.6	77.5
SIMLPE [19]	6.5	-	-	55.8	12.4	-	-	77.3	8.1	-	-	48.6	8.8	-	-	73.8
TFAN [17]	6.5	17.1	43.1	54.9	**12.3**	**28.4**	62.4	**76.9**	7.9	17.5	38.3	48.3	8.6	22.1	55.4	72.0
April-GCN [14]	**6.2**	16.4	40.1	50.8	13.0	32.0	71.7	87.8	**7.4**	**17.3**	38.0	47.8	9.0	23.7	59.5	77.0
Ours	6.6	17.8	43.6	55.5	**12.3**	29.2	62.9	77.8	7.8	17.9	38.4	**48.2**	**7.3**	**16.9**	**34.7**	**45.5**
Motion	**Purchases**				**Sitting**				**Sitting Down**				**Taking Photo**			
Milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
PGBIG [40]	12.5	28.7	60.1	73.3	8.8	19.2	42.4	53.8	13.9	27.9	57.4	71.5	8.4	18.9	42.0	53.3
SPGSN [41]	12.8	28.6	61.0	74.4	9.3	19.4	42.3	53.6	14.2	27.7	56.8	70.7	8.8	18.9	41.5	52.7
EqMotion [18]	11.3	**26.6**	59.8	74.1	8.3	**18.1**	**41.1**	**53.0**	13.2	**26.3**	**56.0**	70.4	8.1	17.8	41.0	52.7
SIMLPE [19]	11.7	-	-	**72.4**	8.6	-	-	55.2	13.6	-	-	70.8	7.8	-	-	50.8
TFAN [17]	11.7	27.6	**59.0**	73.1	8.7	19.2	42.8	54.4	13.6	29.1	57.5	**70.3**	**7.7**	**17.5**	**39.7**	**50.5**
April-GCN [14]	**11.0**	26.8	59.2	73.0	**8.0**	**18.1**	41.7	53.3	**13.0**	26.7	56.5	70.8	7.8	18.0	41.5	52.9
Ours	12.1	29.0	61.5	76.1	8.4	19.0	42.3	53.7	14.0	30.2	59.3	72.2	7.8	17.9	40.3	51.4
Motion	**Waiting**				**Walking Dog**				**Walking Together**				**Average**			
Milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
PGBIG [40]	8.9	20.1	43.6	54.3	18.8	39.3	73.7	86.4	8.7	18.6	34.4	41.0	10.3	22.7	47.4	58.5
SPGSN [41]	9.2	19.8	43.1	54.1	-	-	-	-	8.9	18.2	33.8	40.9	10.4	22.3	47.1	58.3
EqMotion [18]	7.9	17.9	41.5	53.1	**16.6**	**35.9**	72.0	85.7	8.1	17.1	**32.4**	**39.2**	9.4	20.9	46.2	57.9
SIMLPE [19]	7.8	-	-	53.2	18.2	-	-	83.6	8.4	-	-	41.2	9.6	-	-	57.3
TFAN [17]	**7.6**	**17.6**	**40.8**	**51.7**	18.0	38.1	**71.0**	**84.0**	8.4	18.0	35.7	43.7	9.5	21.5	46.0	57.3
April-GCN [14]	7.7	18.4	41.6	52.3	16.7	37.0	71.3	**84.0**	**8.0**	17.6	33.2	40.0	**9.2**	21.4	46.6	57.8
Ours	**7.6**	18.4	42.1	53.5	18.3	38.8	72.7	85.7	7.9	**17.0**	33.0	40.0	9.3	**20.1**	**43.3**	**53.8**

Table 2. Long-term prediction across 15 actions and average prediction results of the H3.6M dataset. The best results are highlighted in bold.

Motion	**Walking**		**Eating**		**Smoking**		**Discussion**		**Directions**		**Greeting**		**Phoning**		**Posing**	
Milliseconds	**560**	**1000**	**560**	**1000**	**560**	**1000**	**560**	**1000**	**560**	**1000**	**560**	**1000**	**560**	**1000**	**560**	**1000**
PGBIG [40]	48.1	56.4	51.1	76.0	46.5	69.5	87.1	118.2	**69.3**	100.4	110.2	143.5	65.9	102.7	106.1	164.8
SPGSN [41]	46.9	**53.6**	49.8	**73.4**	46.7	68.6	-	-	70.1	100.5	-	-	66.7	102.5	-	-
EqMotion [18]	50.7	58.8	50.7	76.2	45.4	67.3	87.1	116.8	69.8	101.2	111.6	144.2	65.8	102.3	111.2	169.6
SIMLPE [19]	**46.8**	55.7	49.6	74.5	47.2	69.3	85.7	116.3	73.1	106.7	**99.8**	**137.5**	66.3	103.3	103.4	168.7
TFAN [17]	53.0	64.4	49.5	76.7	48.6	72.4	84.7	116.3	72.4	107.2	100.9	138.5	66.7	105.5	102.1	166.8
April-GCN [14]	49.4	54.8	50.6	74.6	46.5	68.8	88.0	117.7	70.3	**100.0**	111.0	142.6	**65.5**	**100.3**	109.7	166.4
Ours	**46.8**	60.7	**48.8**	74.0	**42.3**	**65.5**	**66.6**	**106.2**	73.3	104.5	103.0	140.9	66.1	104.5	**72.4**	**137.8**
Motion	**Purchases**		**Sitting**		**Sitting Down**		**Taking Photo**		**Waiting**		**Walking Dog**		**Walking Together**		**Average**	
Milliseconds	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000	560	1000
PGBIG [40]	95.3	133.3	74.4	116.1	96.7	147.8	74.3	118.6	72.2	**103.4**	104.7	139.8	51.9	64.3	76.9	110.3
SPGSN [41]	-	-	75.0	116.2	-	-	75.6	118.2	73.5	103.6	-	-	-	-	77.4	109.6
EqMotion [18]	97.4	136.9	**74.2**	116.0	96.9	148.9	77.7	122.4	73.5	105.8	104.4	142.1	**50.4**	62.0	77.8	111.4
SIMLPE [19]	**93.8**	**132.5**	75.4	114.1	95.7	142.4	**71.0**	**112.8**	71.6	104.6	105.6	141.2	50.8	61.5	75.7	109.4
TFAN [17]	96.1	136.1	75.4	**114.0**	**95.2**	**141.8**	71.2	113.9	**70.4**	105.0	107.1	146.5	55.2	70.3	76.5	111.7
April-GCN [14]	96.8	135.4	75.4	117.2	98.7	149.3	75.9	119.1	72.3	103.6	**103.2**	**137.1**	50.5	**61.1**	77.6	109.9
Ours	98.6	136.9	74.5	115.2	96.9	143.3	72.7	115.2	72.9	105.7	108.2	143.9	51.0	65.5	**72.9**	**108.0**

Table 1 provides a comprehensive overview of the short-term prediction results across various actions on the Human3.6M dataset. Our model achieves notable accuracy, particularly excelling in the 'posing' and 'discussion' actions. For instance, our model records the lowest MPJPE for the 'discussion' action across all time intervals, indicating its precision in predicting subtle and repetitive movements. Similarly, for the 'posing' action, our model surpasses all competitors at shorter intervals, reducing MPJPE by 37.3% and 36.8% on

160 ms and 320 ms intervals, respectively, compared to TFAN [17]. This suggests that our model effectively captures complex, nuanced changes in action dynamics over time.

In terms of long-term predictions, as shown in Table 2, our model consistently outperforms others, achieving the best accuracy for both 560 ms and 1000 ms averages. Notably, when comparing the relative percentage reduction in the 'posing' action, our model reduces MPJPE by 29.8% compared to TFAN [17] at 560 ms. These results underscore the model's ability to maintain high predictive accuracy over extended periods, capturing complex temporal dependencies that other models struggle with. Additionally, our model also delivers competitive results in actions such as 'eating', 'directions', and 'walking together', further validating its generalization capabilities across diverse motion categories.

The data from Tables 1 and 2 confirm that the prediction error of our proposed method tends to decrease with time, which means it has the capacity to fix the generation error, in that our model leverages SkDiff to mitigate this degradation. The 2D-DE, in particular, excels at extracting detailed transition features from generated difference information of the time series. Therefore, it is able to identify unreasonable predictions and reduce their weight, resulting in more precise long-term predictions.

Furthermore, by integrating features from both the temporal and spatial dimensions, our model is adept at finding the correlation between temporal and spatial information, which is crucial for accurate predictions over mid-to-long-term periods (320–1000 ms). Among the state-of-the-art models, ours consistently achieves the highest average accuracy in these mid-to-long-term predictions, demonstrating its strong adaptability and robustness across various temporal scales. For overall evaluation, we consider taking the average MPJPE across all listed temporal scales, thus finding that our model's average MPJPE is 4.7% lower than that of the second-best model April-GCN [14].

4.4.2. Results on 3DPW

Table 3 provides a comprehensive comparison of our model's 3D spatial predictions on the 3DPW dataset, demonstrating its superior performance over existing state-of-the-art approaches across both short-term and long-term prediction intervals. Unlike other models, which often face difficulties in balancing performance across different temporal horizons, our model consistently achieves outstanding results across both short-term and long-term predictions using a unified framework.

A closer analysis of the data presented in Table 3 further highlights the significant advantages of our model in both short-term (200 ms) and long-term (1000 ms) predictions. Our approach reduces the MPJPE by 10.6% at the 200 ms interval and by 31.6% at the 1000 ms interval compared to the second-best model DPnet [42]. The long-term prediction result also proves that our model is able to decrease the generation error owing to 2D-DE.

These results underscore the effectiveness of our model in maintaining high predictive accuracy across varying databases, demonstrating its generalization in capturing complex 3D motion patterns in both man-made and natural scenarios. We also evaluate our model overall with the same method as in the analysis of the performance on Human3.6M, and it turns out that our model has an average MPJPE 26.6% lower than the second-best model DPnet [42].

Table 3. Comparison of prediction results on the 3DPW dataset. The best results are highlighted in bold.

Milliseconds	**200**	**400**	**600**	**800**	**1000**
DMGNN [11]	37.3	67.8	94.5	109.7	123.6
LTD [12]	35.6	67.8	90.6	106.9	117.8
MSR [39]	37.8	71.3	93.9	110.8	121.5
PGBIG [40]	35.3	67.8	89.6	102.6	109.4
DPnet [42]	29.3	58.3	79.8	94.4	104.1
April-GCN [14]	30.4	61.8	88.0	98.2	105.4
Ours	**26.2**	**44.7**	**59.5**	**66.9**	**71.2**

4.5. Visualization Analysis

To further validate our model, we selected four typical poses from the prediction results for visualization. As illustrated in Figure 3, we compared our approach with HisRep [13], used as a baseline. It is clear that our model demonstrates better performance by comparing the ground truth similarity between the predictions yielded by HisRep and our model. Our model proves effective for slight motion changes, as it captures transition features obtained by 2D-DE. In the visualizations shown in Figure 3a,d, it is evident that the baseline tends to generate rapid changes in predictions regardless of the history of motion changing characteristics. In contrast, our model effectively identifies the motion transition features and naturally generates high-quality motions aligned with them.

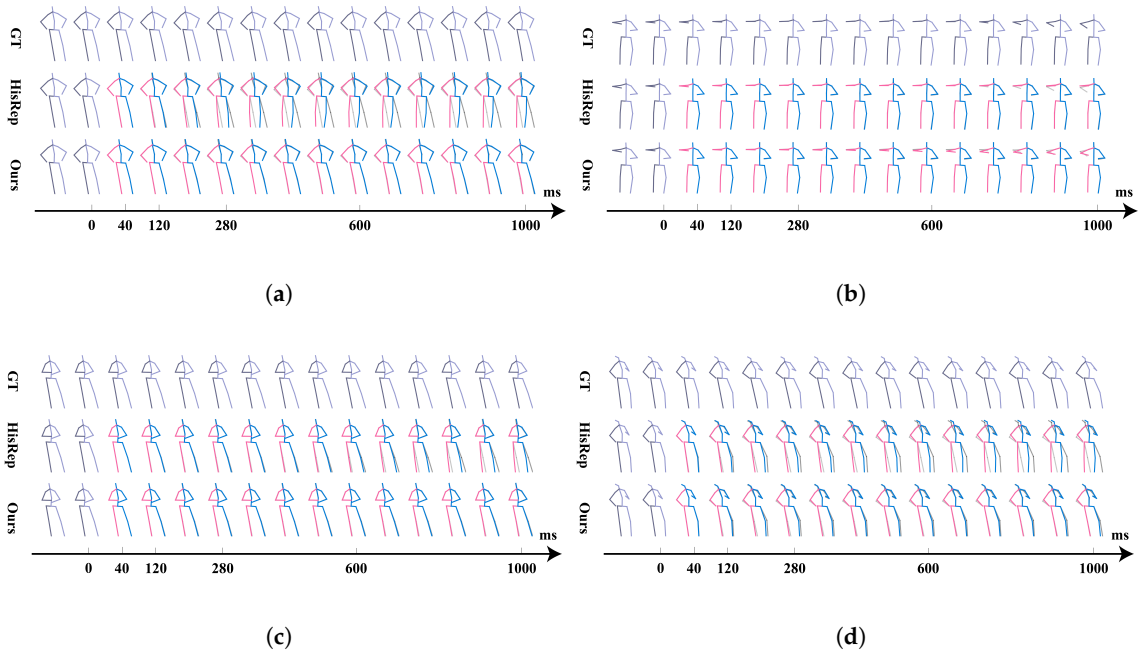

Figure 3. Visualization results of 2DHnet. The visual demonstration includes four common motions: (**a**) motion of 'discussion', (**b**) motion of 'eating', (**c**) motion of 'waiting', and (**d**) motion of 'posing', all from the H3.6M dataset. The labels "GT", "HisRep", and "Ours" represent the ground truth, predictions from the HisRep method, and predictions from our approach, respectively. The predicted motion is highlighted with pink and blue, while the ground truth is colored with gray and lilac. We mix the prediction and ground truth together for better comparison, and it is clear that our method provides more accurate predictions, effectively capturing the nuances in human motion series.

Our method also excels in slight adjustments, as shown in Figure 3b,c. It captures subtle movements, with the skeletons transitioning smoothly and maintaining pace with the motion. It is shown that our model has the capability of correcting the generation error, which could occur during upper or lower body motion prediction.

From the visualization results, it is evident that our model can distinguish between different motion classes, allowing for more tailored responses to historical skeletal sequences. It is highlighted that our model is better at handling the smooth transition of motion, avoiding large differences between reality and prediction.

4.6. Ablation Study

We conduct ablation studies to evaluate the influence of the numbers of GCLs in AST-DE, different configurations, and the SkDiff Extractor latent feature size of our approach on the Human3.6M dataset.

4.6.1. Effect of the Number of GCLs in AST-DE

We separately tested the impact of GCN layers in the AST-DE on model performance when using MLP and Add as the SkDiff Mixer (Figure 4). The data demonstrate that the model achieves high predictive accuracy when the number of GCLs is set to 18. We believe this is because as the number of layers in the model increases, it can extract more spatial features of the human skeleton, which is beneficial for learning structural information. In addition, when using Add as a fusion method, the model performs better than using MLP. We speculate that addition increases the information content of features in the same dimension and adds detailed information.

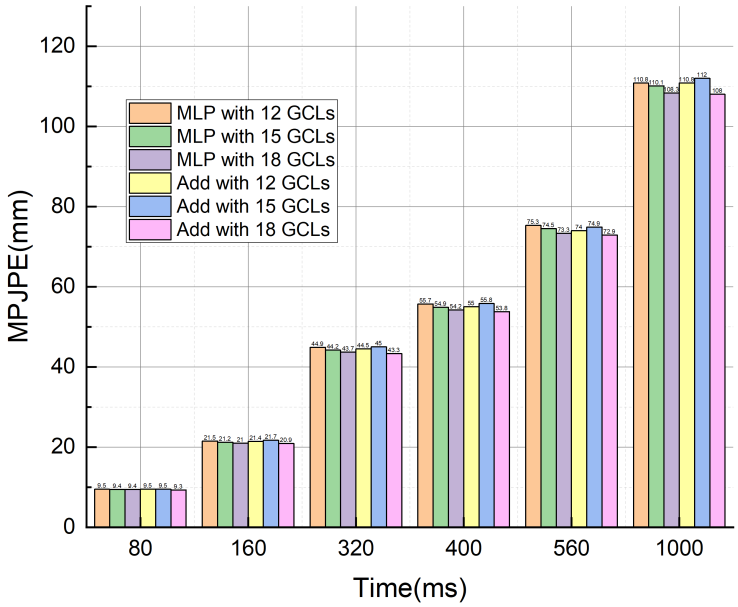

Figure 4. Ablation tests on the number of GCLs in AST-DE.

4.6.2. Effect of Different Configurations

To better understand the impact of various configurations on the model's performance, we conducted several experiments where we varied whether AST-DE and 2D-DE were used, the depth of SkDiff, the padding frame used in SkDiff, and the type of Mixer used for the 2D-DE and AST-DE. Our best results were achieved when both the AST-DE and the 2D-DE were present, the SkDiff depth was set to 1, the last frame was used for padding, and the Mixer was Add. The ablation results in Table 4 demonstrate the impact of different configurations on the model's performance, and our analysis is presented below.

Effect of components: To examine the effect of the components, we first evaluated scenarios where the model used only one of the components. Subsequently, we tested configurations that included both components. Our findings are summarized as follows:

- Enabling only the AST-DE significantly improves model performance, particularly across various time horizons. This suggests that the AST-DE effectively captures the correlation between temporal sequence attention features and spatial graphical features.
- When the AST-DE is absent and the model relies solely on the 2D-DE for prediction, the performance deteriorates markedly. This indicates that the differential information alone causes a loss of skeletal data and fails to model complete and effective motion semantics.
- Integrating the 2D-DE with the AST-DE enhances the model's long-term prediction capabilities. As noted earlier, our approach is effective in correcting long-term errors using the 2D-DE. As shown in Table 4, the combined model outperforms the models using a single component, reducing MPJPE by 1.4 at 1000 ms.

Table 4. Ablation results on the existence of AST-DE or 2D-DE, SkDiff depth, the frame used in SkDiff padding and the Mixer type of 2D-DE and AST-DE. The best results of the ablation group are highlighted in bold. We also evaluate the effect TSAtt existence, which is demonstrated with the symbol in the table.

TSAtt	Depth	Padding	Mixer	80 ms	160 ms	320 ms	400 ms	560 ms	1000 ms
✓	-	-	-	**9.3**	**20.9**	43.6	54.1	73.6	109.7
	1	Last	MLP	362.2	363.1	367.1	369.3	550.2	695.1
✓	1	Last	MLP	9.4	21.0	43.7	54.2	73.3	108.3
✓	2	Last	MLP	9.7	21.8	45.4	56.1	75.4	111.0
✓	3	Last	MLP	9.4	21.2	44.1	54.5	73.4	108.1
✓	1	First	MLP	9.7	21.4	44.2	54.7	74.1	109.5
✓	2	First	MLP	9.5	21.4	44.3	55.0	74.3	109.8
✓	3	First	MLP	9.6	21.4	44.7	55.4	74.7	109.6
✓	1	Last	Add	**9.3**	**20.9**	**43.3**	**53.8**	72.9	108.0
✓	2	Last	Add	9.6	21.6	44.8	55.3	74.3	110.5
✓	3	Last	Add	9.5	21.5	44.3	54.8	73.7	109.7
✓	1	First	Add	9.6	21.6	44.5	55.0	73.9	110.1
✓	2	First	Add	9.5	21.5	44.4	55.0	74.3	111.1
✓	3	First	Add	9.4	21.2	43.7	53.9	**72.8**	**107.2**

Effect of SkDiff depth: As the SkDiff depth increases from 1 to 3, model performance varies across time horizons. Generally, a depth of 1 or 3 yields better results, while a depth of 2 slightly increases errors. This suggests that the 2-SkDiff model is more sensitive to noise, whereas the 3-SkDiff model has a higher capacity to mitigate noise and capture more detailed motion trends for prediction. This is supported by the experimental result where the depth is 3, the padding is "first", and the mix method is "Add".

Effect of padding frame position: The choice of padding frame position ("last" vs. "first") also impacts model performance. As mentioned in Section 3.3, repeating the last frame may enhance the model's ability to eliminate inaccurate predictions, while repeating the first frame emphasizes recent motion dynamics. The results indicate that using the "last" frame for padding generally leads to better performance than using the "first" frame, likely due to increased robustness against noise at the end of the SkDiff block. However, this trend is not always consistent, as the initial weights can also affect the outcome; hence, the "first" frame padding strategy may also yield competitive results.

Effect of SkDiff Mixer type: When comparing the use of MLP (multi-layer perceptron) and Add as the mix method for the two components, we observe that the MLP approach generally achieves higher MPJPE values for shorter time horizons (80 ms to 320 ms) but lower MPJPE errors for longer time horizons (560 ms to 1000 ms). This suggests that MLPs are more effective at learning complex nonlinear relationships but may suffer from overfitting, leading to less accurate long-term predictions. Consequently, we replaced the MLP with Add, which theoretically better compensates for the AST-DE, given their linear relationship. Although the best results were obtained with the Add configuration, we observed that the overall performance difference between the two Mixer types was relatively small. This finding suggests that a simple Add might not be sufficient to extract effective motion semantics.

4.6.3. Effect of the Number of MLP Channels in SKDiff Block

When we chose the Add as SkDiff Mixer to fuse features, we conducted ablation experiments on the number of channels in the MLP (Table 5). The experimental results show that the model's prediction results are the most accurate when the number of channels is 128. We speculate that as the number of channels increases, the model overly focuses on detailed features, resulting in overfitting.

Table 5. Ablation tests on the number of MLP channels in SkDiff block when using Add Mixer. The best results in the ablation study group are highlighted in bold.

Number of Channels	80 ms	160 ms	320 ms	400 ms	560 ms	1000 ms
128	**9.3**	**20.9**	**43.3**	**53.8**	**72.9**	**108.0**
256	9.5	21.4	44.1	54.5	73.4	109.6
512	9.6	21.6	44.7	55.4	74.6	110.4

4.6.4. Effect of the Number of Channels in SkDiff Extractor and SkDiff Mixer

In the 2D-DE, we tested the number of channels C_d for the SkDiff Extractor and SkDiff Mixer, experimenting with C_d = 128, 256, and 512. As can be seen from Figure 5, when C_d = 128, the model achieves the best prediction performance across all time intervals. We believe this could be attributed to a balance between model complexity and computational efficiency. With too many channels, the model might become overly complex, which could result in overfitting to the training data or increased computational cost without a corresponding improvement in predictive accuracy.

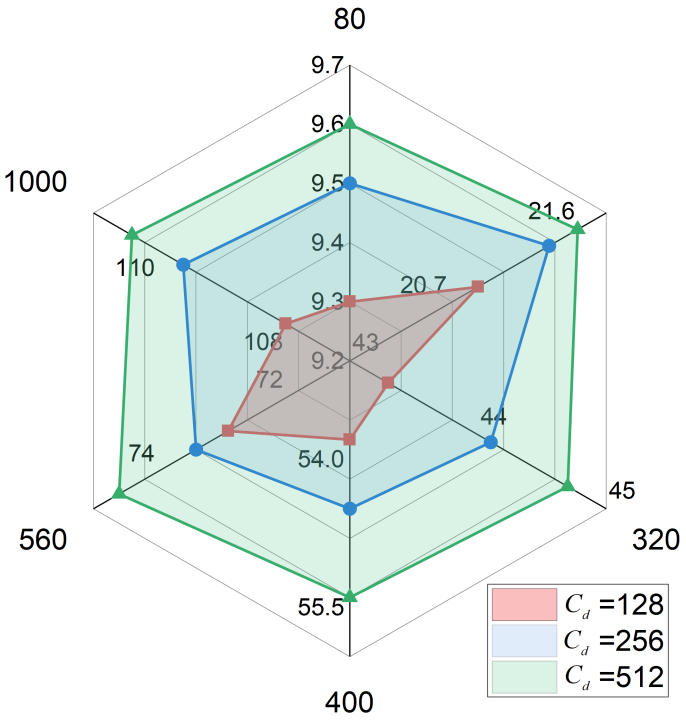

Figure 5. Ablation tests on the number of channels in SkDiff Extractor and SkDiff Mixer.

5. Conclusions

This paper presents Dynamic Differencing-based Hybrid Networks (2DHnet) for improved human motion prediction. Traditional models and recent deep learning approaches often struggle to capture rapid changes in motion dynamics and optimize spatial–temporal features effectively. To address these challenges, 2DHnet introduces two key modules: the Dynamic Differential Dependencies Extractor (2D-DE), which captures dynamic features such as velocity and acceleration; and the Attention-based Spatial–Temporal Dependencies Extractor (AST-DE), which enhances spatial–temporal correlations. Our 2DHnet provides a more comprehensive representation of human motion by combining these modules in a dual-branch architecture. Experimental results on the Human3.6M and 3DPW datasets demonstrate that 2DHnet achieves significant improvements over state-of-the-art methods, with average MPJPE reductions of 4.7% and 26.6%, respectively. Even though this work highlights the importance of motion dynamics learning, 2DHnet may lack transparency in

its decision-making processes. Understanding why the model makes certain predictions can be challenging, which may hinder its acceptance in applications requiring high levels of interpretability. Addressing these limitations in future work could involve simplifying the model architecture and improving data augmentation techniques.

Author Contributions: Methodology, J.Z. and R.J.; software, R.J. and C.L.; data curation, R.J. and C.L.; visualization, C.L. and J.Z.; writing—original draft preparation, R.J. and C.L.; writing—review and editing, J.Z.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets used in this research are available open source at the following links: Human3.6M dataset at http://vision.imar.ro/human3.6m (accessed on 20 September 2024); 3DPW dataset at https://virtualhumans.mpi-inf.mpg.de/3DPW/license.html (accessed on 20 September 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ghafir, I.; Prenosil, V.; Svoboda, J.; Hammoudeh, M. A survey on network security monitoring systems. In Proceedings of the 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), Vienna, Austria, 22–24 August 2016 ; IEEE: Piscataway, NJ, USA, 2016; pp. 77–82.
2. MacKenzie, I.S. *Human-Computer Interaction: An Empirical Research Perspective*; Morgan Kaufmann: Cambridge, MA, USA, 2012.
3. Weinland, D.; Ronfard, R.; Boyer, E. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* **2011**, *115*, 224–241. [CrossRef]
4. Wang, J.M.; Fleet, D.J.; Hertzmann, A. Gaussian process dynamical models for human motion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 283–298. [CrossRef] [PubMed]
5. Brand, M.; Hertzmann, A. Style machines. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 23–28 July 2000; pp. 183–192.
6. Zhong, J.; Cao, W. Geometric algebra-based multiscale encoder-decoder networks for 3D motion prediction. *Appl. Intell.* **2023**, *53*, 26967–26987. [CrossRef]
7. Fragkiadaki, K.; Levine, S.; Felsen, P.; Malik, J. Recurrent Network Models for Human Dynamics. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4346–4354.
8. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 5308–5317.
9. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2891–2900.
10. Shi, J.; Zhong, J.; Cao, W. Multi-semantics Aggregation Network based on the Dynamic-attention Mechanism for 3D Human Motion Prediction. *IEEE Trans. Multimed.* **2023**, *26*, 5194–5206. [CrossRef]
11. Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; Tian, Q. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 214–223.
12. Mao, W.; Liu, M.; Salzmann, M.; Li, H. Learning Trajectory Dependencies for Human Motion Prediction. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Los Alamitos, CA, USA, 27 October–2 November 2019; pp. 9488–9496.
13. Wei, M.; Miaomiao, L.; Mathieu, S. History Repeats Itself: Human Motion Prediction via Motion Attention. In Proceedings of the Europe Conference on Computer Vision ECCV, Online, 23–28 August 2020.
14. Gu, B.; Tang, J.; Ding, R.; Liu, X.; Yin, J.; Zhang, Z. April-GCN: Adjacency Position-velocity Relationship Interaction Learning GCN for Human motion prediction. *Knowl.-Based Syst.* **2024**, *292*, 111613. [CrossRef]
15. Zhong, J.; Cao, W. Geometric algebra-based multiview interaction networks for 3D human motion prediction. *Pattern Recognit.* **2023**, *138*, 109427. [CrossRef]
16. Cao, W.; Li, S.; Zhong, J. A dual attention model based on probabilistically mask for 3D human motion prediction. *Neurocomputing* **2022**, *493*, 106–118. [CrossRef]
17. Du, X.; Wang, Y.; Li, Z.; Yan, S.; Liu, M. TFAN: Twin-Flow Axis Normalization for Human Motion Prediction. *IEEE Signal Process. Lett.* **2024**, *31*, 486–490. [CrossRef]
18. Xu, C.; Tan, R.T.; Tan, Y.; Chen, S.; Wang, Y.G.; Wang, X.; Wang, Y. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1410–1420.

19. Guo, W.; Du, Y.; Shen, X.; Lepetit, V.; Alameda-Pineda, X.; Moreno-Noguer, F. Back to mlp: A simple baseline for human motion prediction. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 4809–4819.
20. 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, Salt Lake City, UT, USA, 18–23 June 2018.
21. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
22. Liu, J.; Wang, G.; Duan, L.Y.; Abdiyeva, K.; Kot, A.C. Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks. *IEEE Trans. Image Process.* **2018**, *27*, 1586–1599. [CrossRef] [PubMed]
23. Tang, J.; Wang, J.; Hu, J.F. Predicting human poses via recurrent attention network. *Vis. Intell.* **2023**, *1*, 18. [CrossRef]
24. Zhong, C.; Hu, L.; Zhang, Z.; Ye, Y.; Xia, S. Spatial-Temporal Gating-Adjacency GCN for Human Motion Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
25. He, Z.; Zhang, L.; Wang, H. An initial prediction and fine-tuning model based on improving GCN for 3D human motion prediction. *Front. Comput. Neurosci.* **2023**, *17*, 1145209.
26. Fu, J.; Yang, F.; Dang, Y.; Liu, X.; Yin, J. Learning Constrained Dynamic Correlations in Spatiotemporal Graphs for Motion Prediction. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 14273–14287. [CrossRef] [PubMed]
27. Martínez-González, A.; Villamizar, M.; Odobez, J.M. Pose Transformers (POTR): Human Motion Prediction with Non-Autoregressive Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
28. Mi, L.; Ding, R.; Zhang, X. Skeleton-based human motion prediction via spatio and position encoding transformer network. In Proceedings of the International Conference on Artificial Intelligence, Virtual Reality, and Visualization (AIVRV 2022), Chongqing, China, 2–4 September 2022; SPIE: Bellingham, WA, USA, 2023; Volume 12588, pp. 186–191.
29. Zhao, M.; Tang, H.; Xie, P.; Dai, S.; Sebe, N.; Wang, W. Bidirectional transformer gan for long-term human motion prediction. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 163 [CrossRef]
30. Meneses, M.; Matos, L.; Prado, B.; Carvalho, A.; Macedo, H. SmartSORT: An MLP-based method for tracking multiple objects in real-time. *J. -Real-Time Image Process.* **2021**, *18*, 913–921. [CrossRef]
31. Cao, G.; Huang, W.; Lan, X.; Zhang, J.; Jiang, D.; Wang, Y. MLP-DINO: Category Modeling and Query Graphing with Deep MLP for Object Detection. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), Jeju, Republic of Korea, 3–9 August 2024.
32. Chen, S.; Xie, E.; Ge, C.; Chen, R.; Liang, D.; Luo, P. Cyclemlp: A mlp-like architecture for dense visual predictions. *IEEE Trans. Pattern Anal. Mach. Intell.* 2021. Volume abs/2107.10224. Available online: https://arxiv.org/abs/2107.10224 (accessed on 6 December 2024).
33. Boughrara, H.; Chtourou, M.; Amar, C.B. MLP neural network based face recognition system using constructive training algorithm. In Proceedings of the International Conference on Multimedia Computing & Systems, Tangiers, Morocco, 10–12 May 2012.
34. Shahreza, H.O.; Hahn, V.K.; Marcel, S. MLP-Hash: Protecting Face Templates via Hashing of Randomized Multi-Layer Perceptron. *arXiv* **2022**, arXiv:2204.11054.
35. Bouazizi, A.; Holzbock, A.; Kressel, U.; Dietmayer, K.; Belagiannis, V. MotionMixer: MLP-based 3D Human Body Pose Forecasting. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 23–29 July 2022; pp. 791–798.
36. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
37. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
38. Von Marcard, T.; Henschel, R.; Black, M.J.; Rosenhahn, B.; Pons-Moll, G. Recovering accurate 3d human pose in the wild using imus and a moving camera. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 601–617.
39. Dang, L.; Nie, Y.; Long, C.; Zhang, Q.; Li, G. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11467–11476.
40. Ma, T.; Nie, Y.; Long, C.; Zhang, Q.; Li, G. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6437–6446.
41. Li, M.; Chen, S.; Zhang, Z.; Xie, L.; Tian, Q.; Zhang, Y. Skeleton-parted graph scattering networks for 3d human motion prediction. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 18–36.
42. Tang, J.; Zhang, J.; Ding, R.; Gu, B.; Yin, J. Collaborative multi-dynamic pattern modeling for human motion prediction. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3689–3700. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.