Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 7064910, 12 pages
https://doi.org/10.1155/2020/7064910

Research Article

A Hierarchical Static-Dynamic Encoder-Decoder Structure for 3D Human Motion Prediction with Residual CNNs

Jin Tang, Jin Liu, and JianQin Yin

School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100096, China

Correspondence should be addressed to JianQin Yin; jqyin@bupt.edu.cn

Received 19 March 2020; Accepted 10 August 2020; Published 27 August 2020

Academic Editor: Ibrahim Zeid

Human motion prediction aims at predicting the future poses according to the motion dynamics given by the sequence of history poses. We present a new hierarchical static-dynamic encoder-decoder structure to predict the human motion with residual CNNs. Specifically, to better mine the law of the motion, a new residual CNN-based structure, v-CMU, is proposed to encode not only the static information but also the dynamic information. Based on v-CMU, a hierarchical structure is proposed to model different correlations between the different given poses and the predicted pose. Moreover, a new loss function combining the static and dynamic information is introduced in the decoder to guide the prediction of the future poses. Our framework features two-folds: (1) more effective dynamics mined due to the fusion of information of the poses and the dynamic information between poses and the hierarchical structure; (2) better decoding or prediction performance, thanks to the mid-level supervision introduced by the new loss function considering both the static and dynamic losses. Extensive experiments show that our algorithm can achieve state-of-the-art performance on the challenging G3D and FNTU datasets. The code is available at https://github.com/liujin0/SDnet.

1. Introduction

3D human motion prediction can be regarded as a problem to predict the future poses according to the spatial correlation and the temporal dynamics mined from the observed poses. The traditional methods are based on the encoder-decoder framework. The encoder is used to mine the motion dynamics, which is used by the decoder to generate the future poses. Obviously, the motion dynamics modeling is the key to predict poses.

To better mine the human dynamics, we first analyze the characteristics of the human motion. The motion often includes the relatively static and moving dynamic parts, for example, for the action sequence of "eating," the hand joint may have large movement and other joints may relatively be steady. At the same time, as stated in [1], the human vision separately models the relatively static and dynamic information. However, most existing methods only use residual connection to introduce the dynamic information, which is a strongly coupled static and dynamic information modeling method. This motivates us to present a new scheme to explicitly predict the static and dynamic poses in a relatively weak coupled way.

For dynamics modeling, recurrent neural networks (RNNs) are usually used [2, 3]. However, it is known that RNN-based methods cannot well model the long-term dynamics, which is important to predict the human motion. Besides, as argued in [4], RNN structures have some other problems, for example, error accumulation and discontinuity results. Consequently, several works have proposed to model human motion by CNNs [5]. In this paper, we also model the static and dynamic information based on CNNs and we present a new hierarchical CNN based encoder-decoder structure, static-dynamic network (SDNet), to predict the future poses, leading to augmented performance than previous ones, as illustrated in Figure 1.

When encoding the observed motion dynamics, a new structure, v-CMU (velocity-cascade multiplicative unit) is presented, and not only the history poses but also the movement information of poses from the consecutive

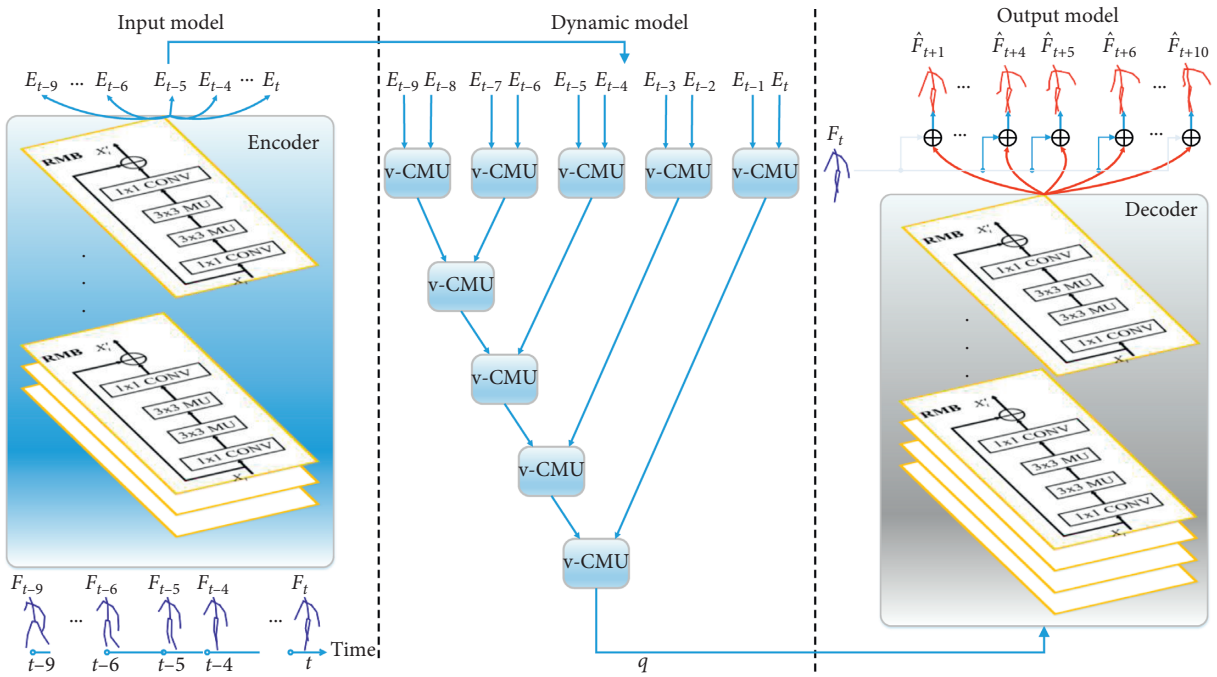


Figure 1: The SDnet architecture for motion prediction, which comprises of the input model, dynamic model, and output model. The encoder of the input model processes spatial appearances and sends the mapped features into the dynamic model. The v-CMUs of the dynamic model explicitly capture the dynamic motion information between the adjacent frames. Then, the decoder maps the features from the dynamic model into the predicted velocity information. Eventually, the output future poses are predicted with the velocity information and the static information from the last input pose.

frames are modeled. After encoding, a hierarchical asymmetric network is employed to model the spatiotemporal information and highlight the dynamics of the historical human poses by using blocked v-CMUs. To keep the features of the last pose, the hierarchical asymmetric structure network models the different contributions of the previously given frames by controlling the number of passing v-CMUs. When decoding the encoded motion dynamics, we introduce a mid-level supervised signal to guide the modeling of the dynamic information. Using the mid-level supervision, on one hand, we can decode the dynamic and static information of the motion in a relatively weak manner; on the other hand, we can introduce different dynamic information for different given poses.

In summary, our contributions are as follows:

(1) We propose a new hierarchical asymmetric structure, SDNet, to predict the future poses. Different from the existed models, SDNet features three branches: SDNet can explicitly model the dynamic motion information in the encoder; SDNet decodes the learned dynamics into the future poses by modeling the static and dynamic information separately; SDNet models the different correlations between different temporal frames and predicted frames using a hierarchical structure.

(2) A new structure, v-CMU, is presented to explicitly model the dynamic motion information. The new proposed v-CMU inputs not only the consecutive frames but also the temporal difference between the consecutive frames; therefore, the v-CMU can explicitly model the dynamic motion information.

(3) A mid-level supervised signal is constructed to guide explicitly modeling the dynamic information of the motion. This mid-level supervision makes our framework to separately model the static and dynamic information and introduce different dynamic information for different predicted frames.

The remainder of the paper is organized as follows. Section 2 investigates the related work. Section 3 discusses our model in detail. The datasets, evaluation criteria, experimental-based comparisons of different methods, and ablation studies are presented in Section 4. Finally, conclusions and future work directions are stated in Section 5.

2. Related Work

Many recurrent neural networks have been designed for predictive learning of spatiotemporal data based on modeling historical frames. Fragkiadaki et al. [6] proposed the encoder-recurrent-decoder (ERD) model for recognition and prediction of human body pose in videos and motion capture, which extended previous long short-term memory (LSTM) models. Martinez et al. [2] further extended this scheme by modeling the velocity of joints instead of directly estimating the body pose and employed a single linear layer for pose feature encoding and hidden state decoding. Tang et al. [7] proposed a new model based on RNN to predict long-term human motions by exploring motion context and

enhancing motion dynamics. For human motion prediction and synthesis, Gopalakrishnan et al. [8] introduced the VTLN-RNN architecture, which used motion derivative features as well as novel multiobjective loss function.

As discussed in [9], conventional recurrent neural networks, such as LSTM [10] and GRU [11], are employed to model motion contexts, which inherently have difficulties in capturing long-term dependencies. Convolutional neural networks (CNNs) have been introduced to solve the problem of motion prediction [12, 13]. Li et al. [14] proposed a convolutional sequence-to-sequence model for human motion prediction based on CNN, which adopted diverse types of convolutional encoders to use both distant and nearby temporal motion information. Van den Oord et al. [15] presented the WaveNet which is built upon causal convolution structures, and this structure could also be used in motion prediction. Liu et al. [16, 17] proposed the SSnet to model the motion dynamics by sliding window over the temporal axis based on dilated convolutional network. Differently, it focused on the observed part of the ongoing action in the untrimmed radios which can include multiple actions. Kalchbrenner et al. [18] proposed the multiplicative unit (MU) which is a nonrecurrent convolutional structure whose neuron connectivity is quite like LSTMs and proposed a residual multiplicative block (RMB) to ease gradient propagation. Xu et al. [19] introduced an entirely CNN-based architecture, PredCNN, to model the spatial information of each frame and capture the temporal evolution of previous frames hierarchically by cascade multiplicative unit (CMU) that receives two consecutive frames as input. Liu et al. [20] applied the PredCNN with the new skeletal representation to get more accurate motion prediction.

3. Method

3.1. Framework of the Model. We propose a novel and end-to-end model for human pose prediction. Our model explicitly captures the temporal dependencies between adjacent frames and predicts all future frames only in one step that can avoid error accumulation.

The framework of the model consists of three parts as shown in Figure 1.

Input model: the input model is used to model the spatial information of each skeleton. As known, the skeleton sequence is a set of joint coordinates. We transform the skeletons' joint coordinates to a pseudo image. Then, the encoders are employed to extract the spatial features of the observed skeletons.

Dynamic model: the dynamic model is a v-CMU hierarchical asymmetric network, which is to model the spatiotemporal information and highlight the dynamics of the historical human poses. On the one hand, we propose v-CMU to explicitly model the difference between two consecutive input frames, which is beneficial to modeling dynamic evolution. On the other hand, the hierarchical asymmetric structure keeps the features of the last pose by passing less v-CMUs. This architecture is enlightened by the key idea that repeating the last body pose gave a relatively small error in the measurement of the Euclidean distance between the ground-truth [2, 7].

Output model: the output model is the static and dynamic integrated residual module in which the static and dynamic information is integrated from two branches. On one branch of this module, the latest frame is used to retain the static information. On the other branch, the decoder is expected to predict the dynamic information, i.e., velocity. Finally, the future poses are predicted by integrating the static and dynamic information at the fine-grained level.

3.2. Input Model. We select and reorder the 18 joints which are informative enough to represent human motion as shown in Figure 2. Given the skeleton of a person in frame i, the input frame F_i is the observed skeletons $\{F_1, F_2, \ldots, F_t\}$. F_i is an input of size $N \times d$, where N corresponds the number of main joints in each frame, N is 18, and d corresponds the number of dimensions describing each joint (if the describing is only the coordinate, d is 3). F_i can be formulated as follows:

$$F_i = \begin{bmatrix} X_1 & Y_1 & Z_1 \\ X_2 & Y_2 & Z_2 \\ \vdots & \vdots & \vdots \\ X_N & Y_N & Z_N \end{bmatrix}, \quad (1)$$

where each row vector of the matrix F_i is 3D joint coordinate. Then, the input F_i is transformed into the pseudo image.

Because the residual multiplicative block (RMB) [18] has power for spatial modeling with its LSTM-like structure, the RMB is the basic unit and is cascaded by l_e layers as the encoder, where l_e is a hyperparameter. After encoding, each pseudo image is mapped into a feature space to describe spatial information.

3.3. Dynamic Model

3.3.1. v-CMU. As we have mentioned before, motion dynamics modeling is the key to predict poses. In this paper, we design a novel v-CMU based on a new residual CNN to explicitly capture the dynamic motion information between the adjacent frames. Our key motivations of v-CMU are twofolds.

Most existing CNN methods only encode the static information. To encode the static and dynamic information in a relatively weak coupled way, we design a dynamic v-CMU unit based on CMU. The motions of three directions are introduced by calculating the difference between adjacent frames elementwise to focus on predicting dynamic evolution.

For dynamics modeling, RNNs are usually used to capture the *underlying* temporal dependencies in the sequential data. However, to *explicitly* capture the temporal dependencies, the new proposed v-CMU is formulated like a

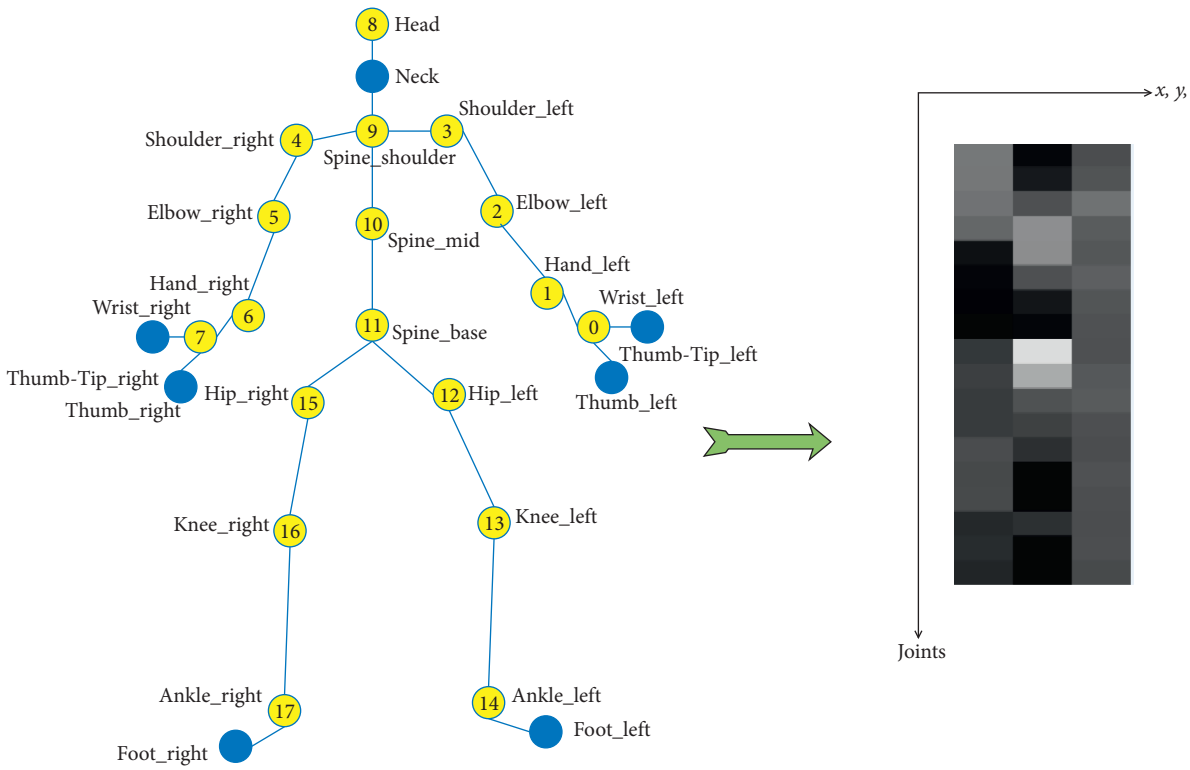


Figure 2: The representation of skeletal data. We select the yellow-colored joints to represent human motion and transform the 3D joint coordinate to pseudo image.

learning residual function to concern the temporal difference between the adjacent frames.

The diagram of the proposed v-CMU is shown in Figure 3, which accepts two consecutive inputs E_{t-1} and E_t generates an output H_t. E_t represents an encoded input image at the current time t, E_{t-1} represents an encoded input image at previous time $t-1$. We first apply CMU to generate a new state which contains rich spatial and hidden temporal features. Then, the difference between the consecutive frames is used to directly add to the output of the CMU elementwisely. By having such a residual structure, we explicitly model the dynamic motion information. Given that "θ" is the parameter of the CMU, the proposed v-CMU can be formulated as follows:

$$H_t = \mathrm{CMU}(E_{t-1}, E_t; \theta) \oplus (E_t - E_{t-1}), \tag{2}$$

where $\mathrm{CMU}(E_{t-1}, E_t; \theta)$ represents the formulation of CMU and $(E_t - E_{t-1})$ is regarded as the residual. The proposed v-CMU has the same number of parameters as CMU, while it is more powerful in modeling dynamic evolutions.

3.3.2. Hierarchical Asymmetric Structure. During all input sequential frames, the last body pose and movement may provide more dependencies for the prediction [2, 7]. Motivated by this idea, we propose a hierarchical asymmetric structure using the proposed v-CMU unit as building blocks. By taking the advantages of this hierarchical asymmetric v-CMU blocks, the latest temporal receptive field is enlarged most explicitly.

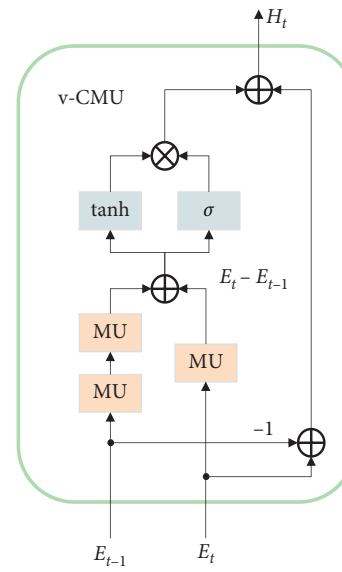


Figure 3: Schematic structure of v-CMU, which adds difference between the consecutive frames to the new state from CMU to explicitly capture the temporal dependencies. MU blocks represent multiplicative units, while tanh and σ blocks denote gated structures containing convolutions along with nonlinear activation functions, and circles represent elementwise operations.

As shown in Figure 4(a), the hierarchical asymmetric network consists of v-CMU blocks and is employed to model the spatiotemporal information and highlights the dynamics

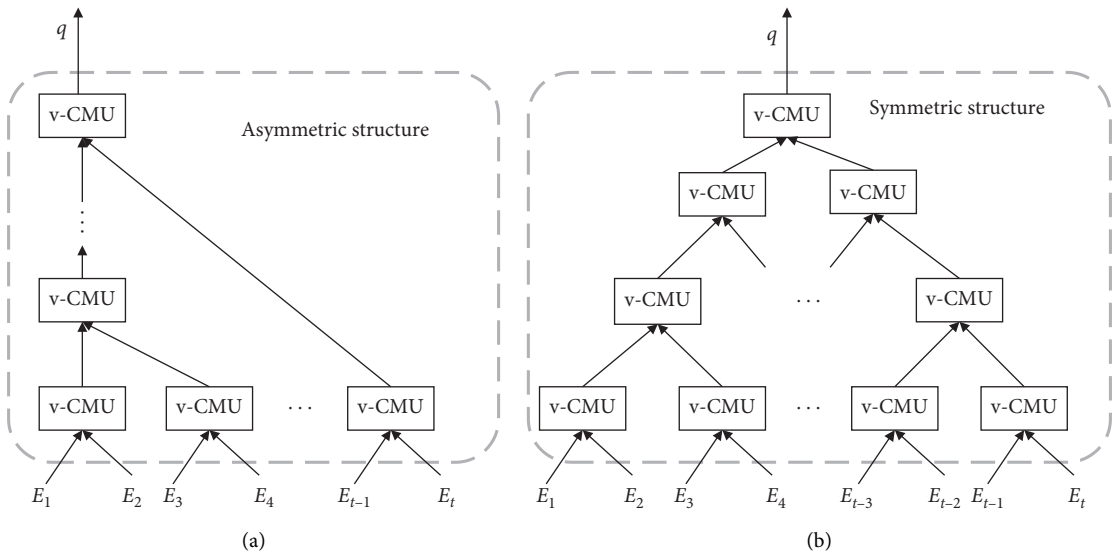


Figure 4: The hierarchical structure based on v-CMUs. (a) Our hierarchical asymmetric structure; the encoded latest frame only passes through 2 v-CMU blocks. (b) Hierarchical symmetric structure, introduced by Liu et al. [20]. The encoded latest frame passes through 4 v-CMU blocks.

of the historical human poses. Comparing with PredCNN [19] which only predicts one future frame, the proposed hierarchical asymmetric structure is a feed-forward architecture and predicts all future frames (10 frames) only in one step that can avoid error accumulation.

Since repeating the last body pose gave a relatively small error, the proposed asymmetric structure can enhance motion dynamics and consider the different contributions of each given frame. The later frames are more correlated with the future frames. As shown in Figure 4(b), each given frame is equally processed in spatiotemporal information [20]. To predict the future pose, the encoded latest frame E_t passes through 4 v-CMU blocks in the symmetric network. On the contrary, our asymmetric network structure handles the different correlations between different given frames and future frames. The encoded latest frame E_t only passes through 2 v-CMU blocks in the asymmetric network to keep the features of the last pose. Besides, our network structure can reduce operations by passing through fewer v-CMU blocks.

3.4. Output Model. To predict the static and dynamic poses in a relatively weak coupled way, a new loss function combining the static and dynamic information is introduced in the output model to guide the prediction of the future poses. As shown in Figure 5, the output model is composed of two branches. One branch is expected to get the static subpose, and the other is used for the dynamic information. We merge the static information with the dynamic information of the two branches to predict the future pose:

$$\widehat{F} = S + \widehat{V}. \tag{3}$$

In equation (3), $\widehat{F}$ represents the future pose, S represents the static information, and $\widehat{V}$ is the dynamic information, and they are calculated from the left and right branches separately.

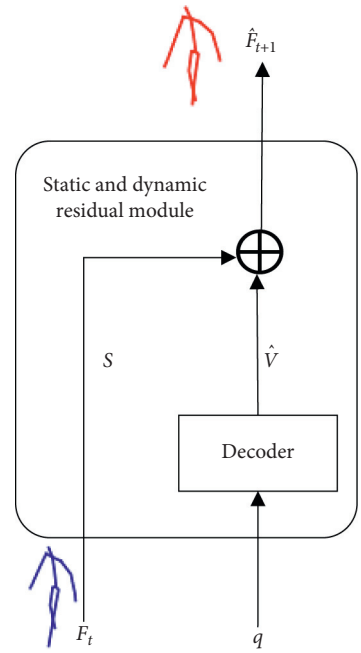


Figure 5: The first future pose that comes from the combination of the static information and the dynamic information. As the static information S, it represents the joint position which comes from input image F_t at time t. $\widehat{V}$ is the dynamic information, which includes the three one-dimensional velocity components of each joint.

3.4.1. Decoding Dynamics under Supervision. The velocity of a moving joint in space can be decomposed orthogonally into three components, and the three velocity components may be different. Therefore, instead of treating the motion of a joint as an indivisible whole, we decompose the three-dimensional motion into three one-dimensional motions

and distinguishing the motion differences at the fine-grained level. Turning now to the modeling of dynamic information, specifically, the three one-dimensional velocity components of each joint are predicted as shown in the right branch in Figure 5. The decoder is similar to the encoder, and they all consist of the basic block RMB, but the number of stacked RMB is l_d in each decoder, where l_d is a hyperparameter. We use the decoder to generate velocity $\widehat{V}$ from the output of the v-CMU hierarchical asymmetric network q that contains rich static and dynamic information.

A new loss function combining the static and dynamic information is proposed. A velocity loss l_v is given in the loss function to learning the dynamic information. The loss function is formulated as equation (4):

$$l = \|F - \widehat{F}\| + l_v. \tag{4}$$

To better guide the learning of velocity during training, a mid-level supervised signal l_v is constructed to assist training. We introduce the supervision information:

$$l_v = \lambda \|V - \widehat{V}\|. \tag{5}$$

where V is the ground-truth velocity of the frame to be predicted relative to the last observed frame:

$$V = F - F_t. \tag{6}$$

This mid-level supervision makes our framework can separately model the static and dynamic information and introduce different dynamic information for different predicted frames.

4. Experiments

4.1. Datasets. We evaluate our proposed model on two datasets, including G3D [21] and the filter NTU RGB + D (FNTU).

G3D: G3D [21] dataset contains 10 subjects performing 20 gaming actions in 7 action sequences captured with Microsoft Kinect. Most sequences contain multiple actions. It consists of 210 samples in total. For a fair comparison, we adopt the same training and test splits as in [20], provided in the released data.

FNTU: FNTU dataset [20] is collected from NTU RGB + D dataset. NTU RGB + D dataset [22] contains 60 action classes and 56,880 video samples. Each video sample contains one action. Based on [20], the FNTU dataset consists of 18102 forward skeleton samples that are selected by removing mutual actions from NTU RGB + D dataset. We follow the same training and test sets, which are made publicly available.

4.2. Metrics and Baselines

Metrics: we follow the standard evaluation protocol used in [19, 20], and choose the mean squared error (MSE) and the mean absolute error (MAE) between the predicted frames and the ground-truth frames in the joint coordinate space as two basic evaluation metrics. As mentioned in [4], angles are not a good representation to evaluate motion prediction. We employ the measurement of the Euclidean distance between the ground-truth pose and our predicted pose in the 3D coordinate space as the error metric.

Baseline: the pose prediction based on 3D joint coordinate sequences is rarely researched. Three baselines are selected to compare with our method. Specific introductions are as follows:

PredCNN: Predictive Learning with Cascade Convolutions [19], an efficient and effective recurrent model for video prediction

S-TE: Symmetric Temporal Encoder [23] converts the mocap frame into the joint coordinate frame in Cartesian coordinates

PISEP^2: Pseudo Image Sequence Evolution-based 3D Pose Prediction [20], the state-of-the-art performance with the new skeletal representation

We use the released code and data by [20] to reproduce the above baselines, and our model is evaluated on the same datasets. The result will be shown later.

4.3. Comparison with Baselines. In this section, we compare our model to the above baselines. In the experiments, we are given 10 frames to predict the future 10 frames for all datasets. To be consistent with the literature, our model has 4 RMBs in the encoder and 6 RMBs in the decoder. Note that, to avoid overfitting on G3D dataset, the number of RMBs in the encoder and the decoder is reduced to 2 and 3. Our results successfully demonstrate the state-of-the-art performance being achieved across all two datasets.

4.3.1. Quantitative Analysis of the Experimental Results. We compare our model to baselines. We show quantitative prediction errors in Table 1. The PredCNN [19] only can predict one frame one time, and the recursive structure causes error accumulation, which leads to the poorest performance. As shown in Table 1, our model achieves state-of-the-art performance on both two datasets. The MSE decreases from 0.1199 to 0.1106 on G3D and 0.1210 to 0.1131 on FNTU. The MAE decreases from 1.1101 to 0.9782 on G3D and 1.1651 to 1.0675 on FNTU. PISEP^2 considers each given frame equally at spatiotemporal information processing in contrast to our v-CMU hierarchical asymmetric network that handles the given latest frame as the most relevant to future frames. On the other hand, our model employs a dynamic decoder to decode the dynamic and static information of the motion in a relatively weak manner. The dynamic decoder can help to better capture dynamic information.

4.3.2. Quantitative Analysis of Framewise Results. The framewise performance of different methods is shown in Figure 6 to analyze the performance of each time-step, where the horizontal axis represents frames and the vertical axis represents MSE or MAE of each frame. The mean MSE of

Table 1: Compare with baselines.

Model	G3D		FNTU	
	MSE	MAE	MSE	MAE
PredCNN [19]	0.1882	1.5713	0.1665	1.6394
S-TE [23]	0.1407	1.2095	0.1425	1.3094
PISEP^2 [20]	0.1199	1.1101	0.1210	1.1651
SDnet	0.1106	0.9782	0.1131	1.0675

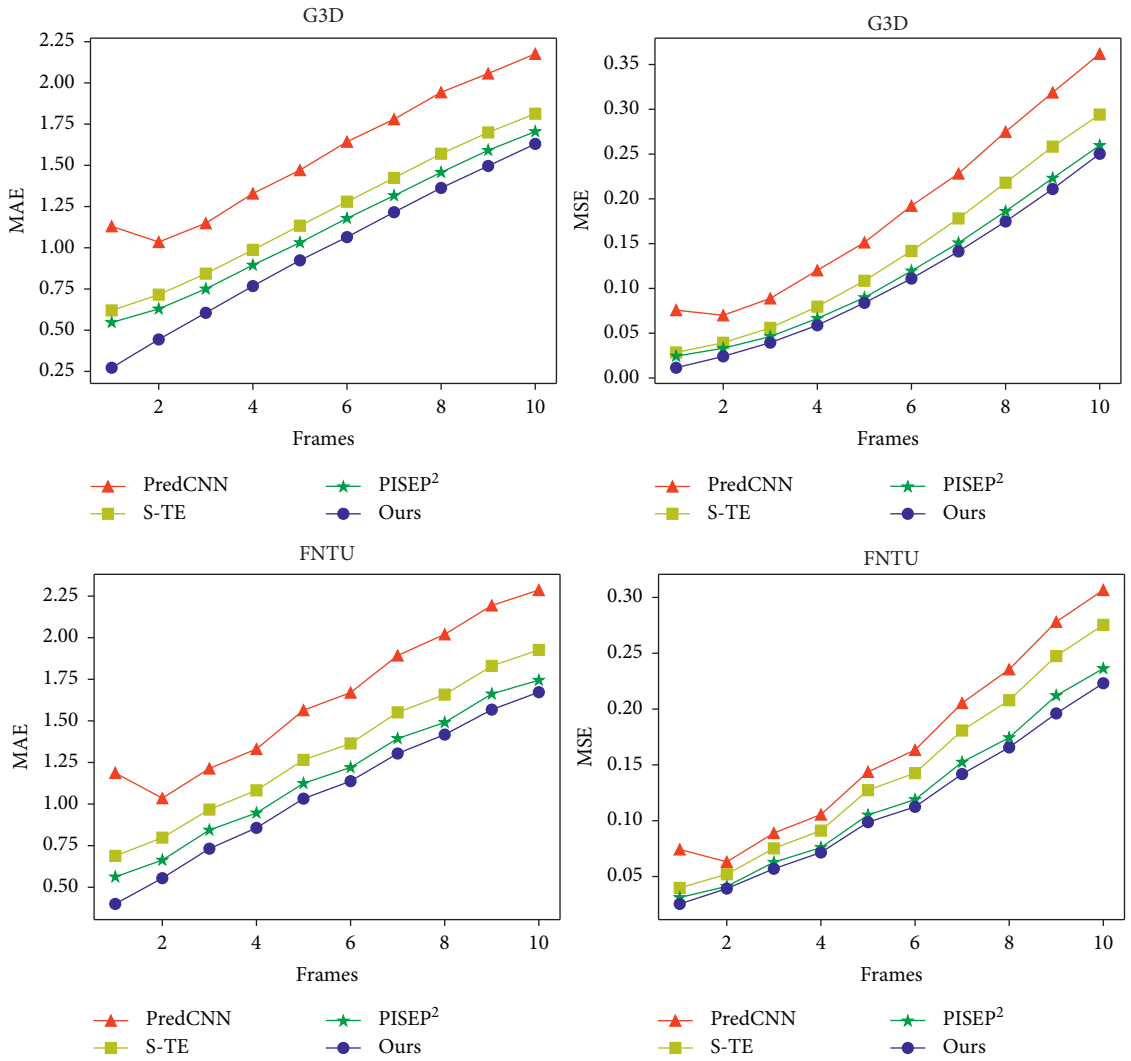
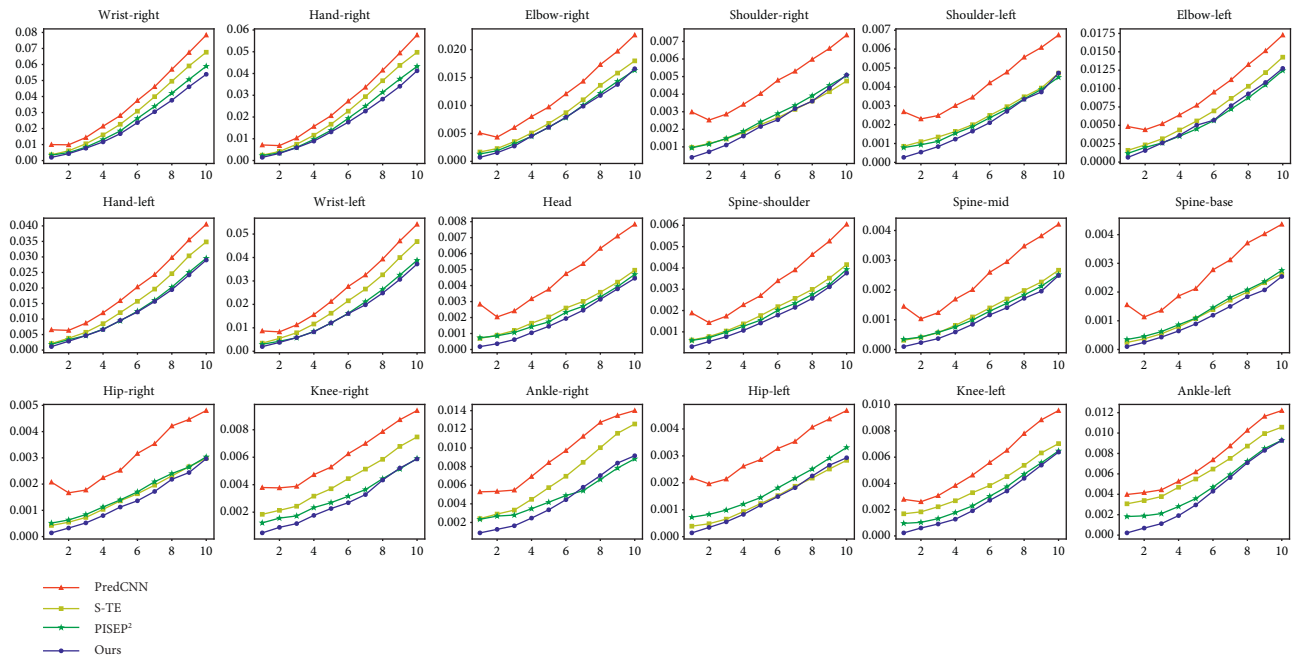


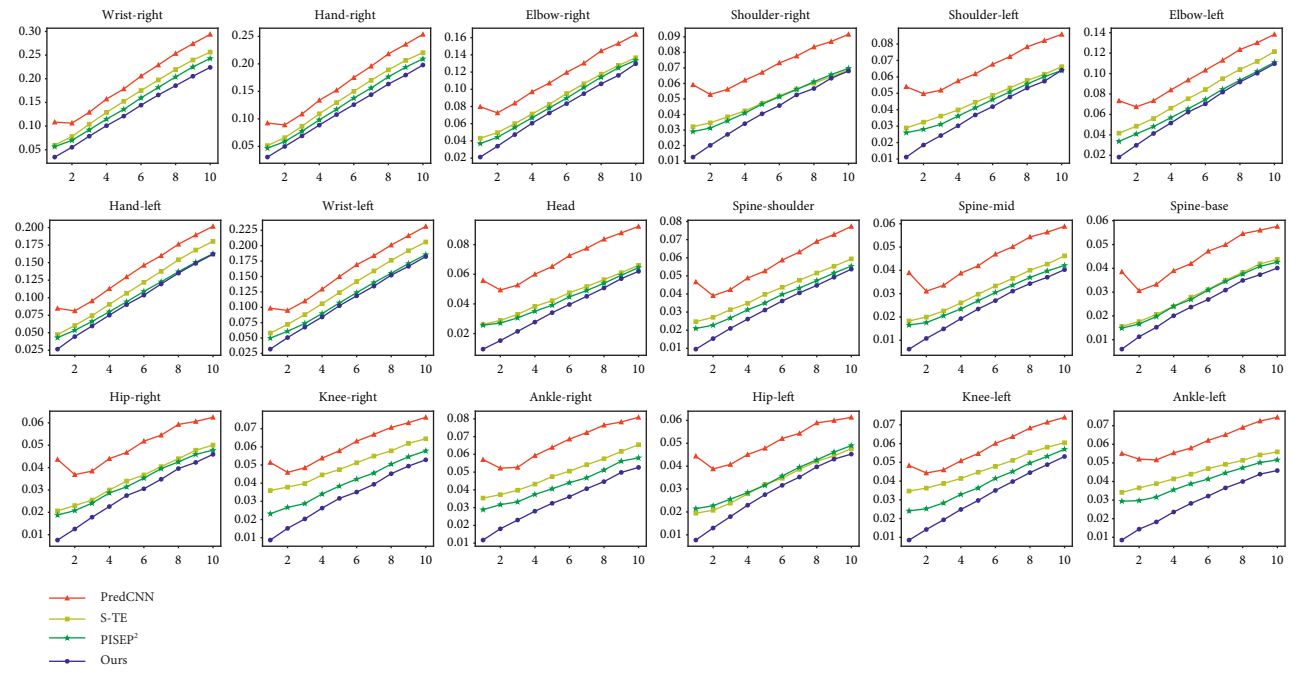
Figure 6: Framewise performance of different methods.

predictive poses for each frame is 0.1106 and 0.1131, and the mean MAE of predictive poses for each pose is 0.9782 and 1.0675 on G3D and FNTU, respectively. The experimental results show that PredCNN [19] suffers from error accumulation that leads to the poorest performance. Compared with the state-of-the-art PISEP^2 [20], our method significantly decreases error at all time-steps, especially for the short-term prediction. However, our method effectively solves the discontinuities in prediction, especially the first prediction frame. This may be due to the coherence of human movements. Paying attention to the last body pose can better predict the following movements since we are repeating the last body pose. As shown in Figure 6, SDnet can significantly enhance the predictive performance on both short-term and long-term predictions. And, our framework achieves the best performance, which further evidence the effectiveness of our proposed method.

4.3.3. Quantitative Analysis of Jointwise Results. To further analyze the performance of our method, the jointwise performance of different methods of each joint is shown in Figure 7, where the horizontal axis represents frames and the vertical axis represents MSE or MAE of each of the joint. (1) On

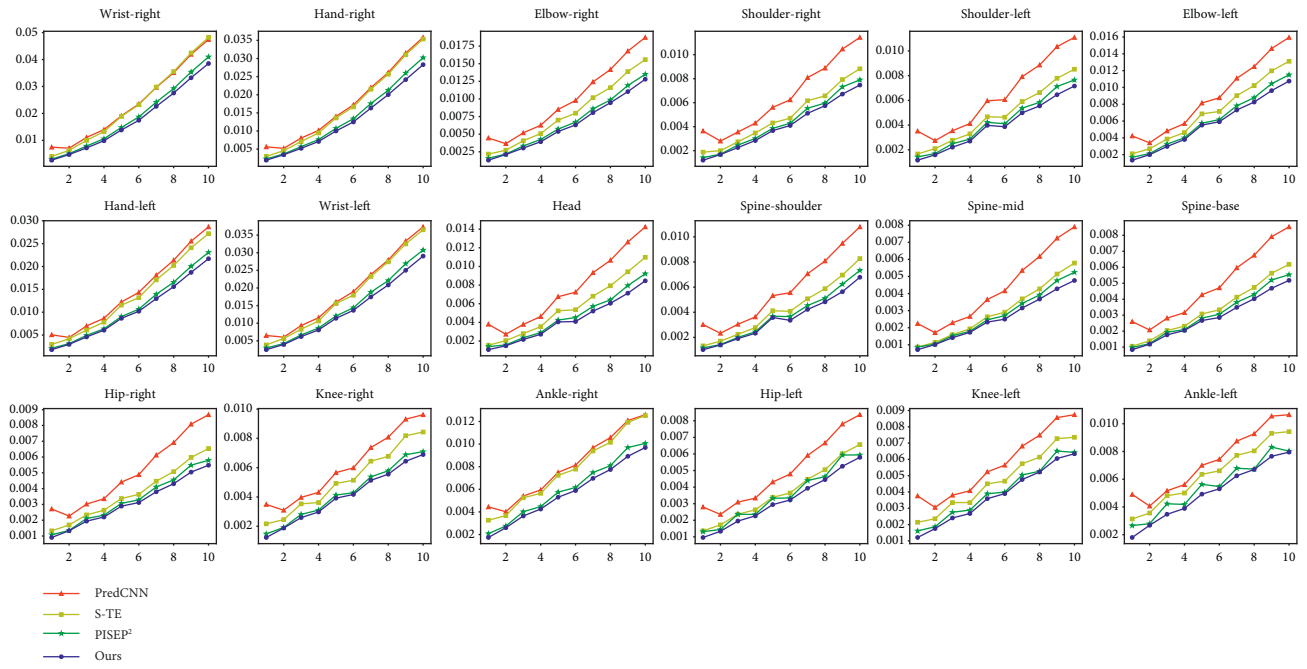


(a)

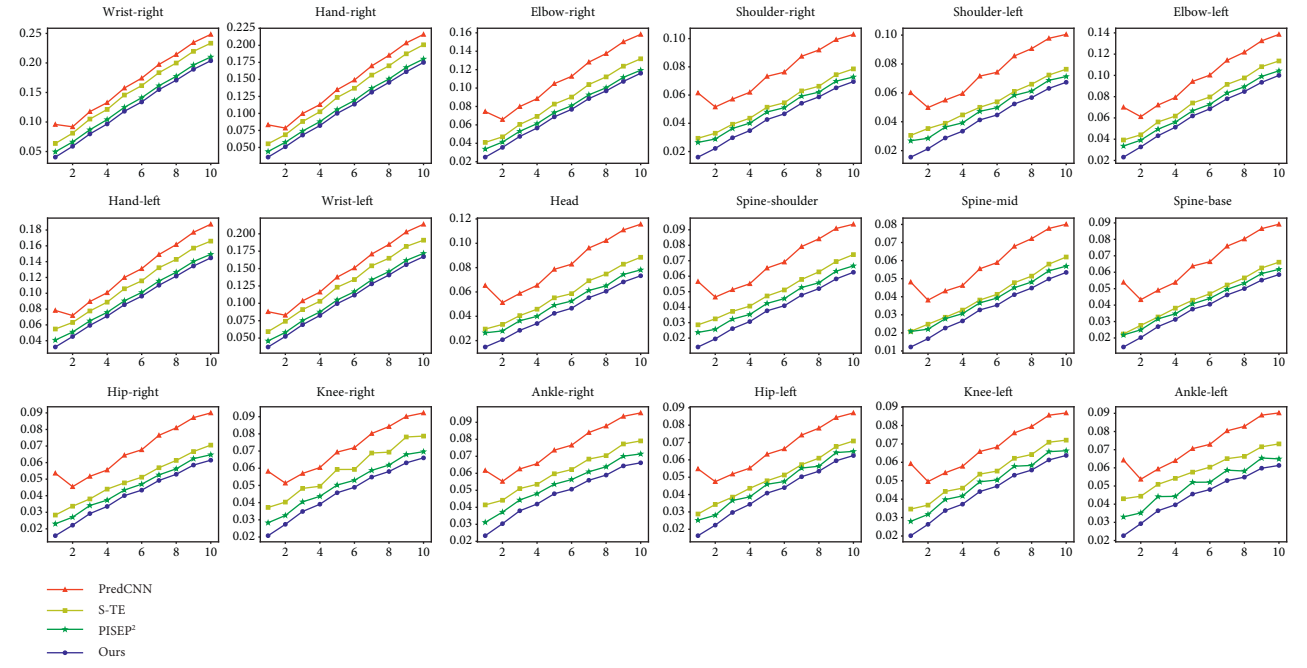


(b)

Figure 7: Continued.



(c)



(d)

FIGURE 7: Jointwise performance of different methods. (a) Jointwise MSE and (b) Jointwise MAE of different methods on G3D. (c) Jointwise MSE and (d) Jointwise MAE of different methods on FNTU.

G3D: in general, the errors of joints of upper limbs are relatively large, especially for the "wrist right," "wrist left," "hand right," and "hand left" on both MSE and MAE. The probable reason for this phenomenon is that the actions on G3D are the upper limbs related actions, and these joints are the most active. Therefore, this may lead to a large error in these joints. Compared with the upper limb joints, the joints of lower limbs or trunk are relatively stable. However, it is interesting that the errors on MAE of the lower limbs and the trunk are approximately close, while on MSE, the errors of lower limbs are larger especially for the "ankle right" and "ankle left." This may be due to MSE being sensitive to action amplitude. Compared with state-of-the-art PISEP^2 [20], the performance of our method significantly extends their performance at all joints especially the first prediction frame, which demonstrates that SDnet can avoid the discontinuities in prediction. (2) On

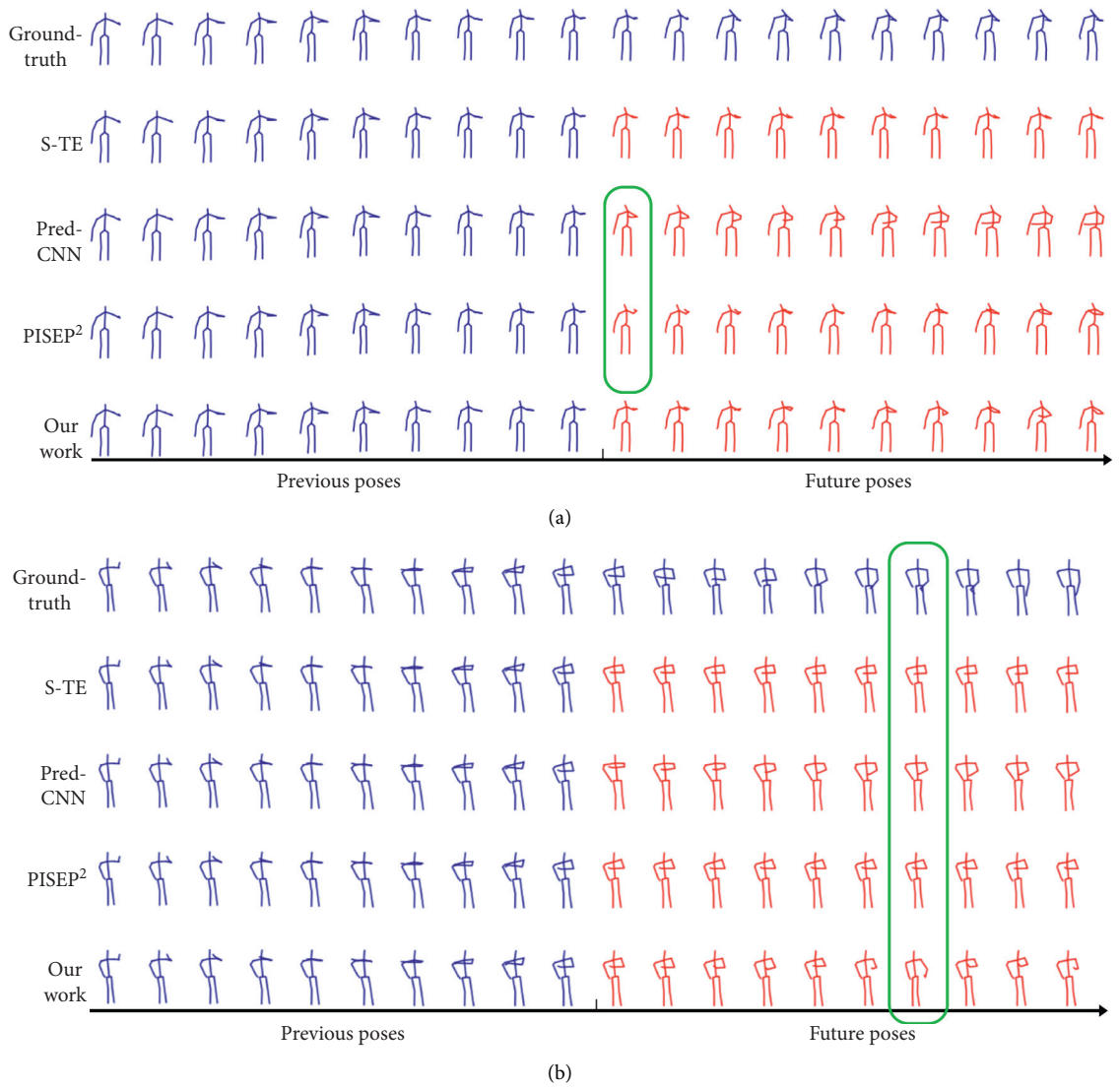


Figure 8: Visualization of predictive performance on (a) G3D and (b) FNTU.

FNTU: the errors of the joints of the upper limbs are relatively large, and the errors of the joints of the lower limbs or trunk are relatively small, which shows comparable results on G3D. This may be due to intense joint movements of the upper limbs. Compared with other methods, our method achieves the best results again for both MSE and MAE. More specifically, compared with [20], our method outperforms at all joints overall, which further shows the effectiveness of our proposed method to capture dynamic information of the previous poses.

4.3.4. Qualitative Analysis of the Experimental Results. We visualize the corresponding framewise prediction results in Figure 8. The previous pose and the ground-truth sequences are shown in blue, and the predicted motion are shown in red. Starting from the top, the five sequences correspond to ground-truth, S-TE, PredCNN, PISEP^2, and our work respectively. However, it seems that the results generated by our model are more accurate and reasonable. As shown in Figure 8(a), PredCNN performs the worst. This may be due to the fact that the first frame of future poses is inaccurate, and the error accumulates. However, S-TE [23] converges to the mean body pose in prediction poses. It is important to note the difference of our method compared with the previous state-of-the-art PISEP^2. We observe that PISEP^2 is worse to stay consistent with the ground-truth than ours especially on the first predicted future frame, as shown in the green circle. Not only PISEP^2 but also some works mentioned in literature [24] are often observed that there is a *significant discontinuity* especially *the first predicted frame*. Our SDnet pays more attention to the motionless directions based on the last body pose. Since we repeat the last body pose, it gives a relatively small error. Noted that even PISEP^2 can avoid the error accumulation, it still worse than ours in the last future frame. The reason can be attributed to that the v-CMU can efficiently extract movement trends, propagate information, and benefit modeling dynamic evolutions.

Table 2: Influence of the v-CMU units on both datasets.

Model	G3D		FNTU	
	MSE	MAE	MSE	MAE
PISEP^2 [20]	0.1199	1.1101	0.1210	1.1651
SDnet	0.1106	0.9782	0.1131	1.0675
Remove v-CMU	0.1236	1.0216	0.1209	1.1143

Table 3: Influence of the asymmetric structure on both datasets.

Model	G3D		FNTU	
	MSE	MAE	MSE	MAE
PISEP^2 [20]	0.1199	1.1101	0.1210	1.1651
SDnet	0.1106	0.9782	0.1131	1.0675
Symmetric structure	0.1147	0.9758	0.1231	1.1226

We provide a qualitative visualization of framewise prediction on FNTU dataset in Figure 8(b). The experimental results show that our framework can avoid error accumulations. Again, our predictions are closer to the ground-truth than that of the baselines. As shown in the green circle, our model predicts motion more *accurately* than other models when the people have a *big move*. Our model captures the spatial information by an LSTM-like block, and the temporal information with a hierarchical asymmetric structure, which consider the different contribution of the previously given frames. Besides, the results generated by other methods converge to the mean body pose. On the contrary, our method can predict motion more reasonably with high dynamics.

4.4. Ablation Studies. To provide a deeper understanding of our method, we next run some ablation studies to evaluate the influence of its components. We use the CMU units instead of v-CMU units to study the contributions of v-CMU unit in our approach using the same hierarchical asymmetric network. To this end, we compare our approach with a symmetric structure network as shown in Figure 4(b) to study the influence of hierarchical asymmetric structure.

The results of these experiments are provided in Tables 2 and 3. As shown in Table 2, the results increase on G3D and FNTU especially the result of MAE when the CMUs replace the v-CMUs. These results show that using our v-CMU units provides a significant boost in accuracy.

Finally, we evaluate the importance of using asymmetric structure vs symmetric structure networks. The results of these experiments, provided in Table 3, demonstrate the benefits of using an asymmetric structure. Note that probably because the G3D dataset is relatively small, the result of symmetric structure on MAE works well. But, the hierarchical asymmetric structure network still gets better results on MSE and MAE of FNTU and MSE of G3D dataset.

Altogether, this ablation study evidences the importance of both aspects of our contribution using the v-CMU to explicitly model the dynamic motion information and hierarchical asymmetric structure to model the different correlations between different temporal frames and predicted frames [24].

5. Conclusions

This paper presents SDnet, a hierarchical convolutional encoder-decoder architecture for static and dynamic pose predictive learning. Specifically, we introduce a velocity-cascade multiplicative unit based on a new residual CNN to explicitly capture the dynamic motion information between the adjacent frames. A hierarchical asymmetric structure using the v-CMUs is proposed to predict all future frames in one step, which enhances motion dynamics and models the different contributions of previously given frames. The proposed SDnet model achieves state-of-the-art performance on the G3D and FTNU datasets. Our future work will focus on the more accurate weight information of the history human body to SDNet to reach a more accurate prediction.

Data Availability

The code to support the findings of this study is available at https://github.com/liujin0/SDnet.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

J. Tang and J. Liu contributed equally.

Acknowledgments

The research in this paper used the NTU RGB + D Action Recognition Dataset made available by the ROSE Lab at the Nanyang Technological University, Singapore.

References

[1] M. Wischnewski, A. Belardinelli, W. X. Schneider, and J. J. Steil, "Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention," *Cognitive Computation*, vol. 2, no. 4, pp. 326–343, 2010.

[2] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2891–2900, Honolulu, HI, USA, July 2017.

[3] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. Moura, "Adversarial geometry-aware human motion prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 786–803, Munich, Germany, May 2018.

[4] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Seoul, South Korea, November 2019.

[5] V. Vukoti, S.-L. Pintea, C. Raymond, G. Gravier, and J. C. van Gemert, "One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network," in *Proceedings of the International Conference on Image Analysis and Processing*, pp. 140–151, Catania, Italy, September 2017.

[6] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Seoul, South Korea, November 2019.

[7] Y. Tang, L. Ma, W. Liu, and W. Zheng, "Long-term human motion prediction by modeling motion context and enhancing motion dynamic," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCA)*, Stockholm, Sweden, July 2018.

[8] A. Gopalakrishnan, A. Mali, D. Kifer, C. L. Giles, and A. G. Ororbia, "A neural temporal model for human motion prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019.

[9] Z. Liu, S. Wu, S. Jin et al., "Towards natural and accurate future motion prediction of humans and animals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] K. Cho, B. Van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: encoder decoder approaches," 2014, https://arxiv.org/abs/1409.1259.

[12] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 1655–1661, San Francisco, CA, USA, February 2017.

[13] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Juan, PR, USA, May 2016.

[14] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5226–5234, Salt Lake City, UT, USA, June 2018.

[15] A. Van den Oord, S. Dieleman, H. Zeny et al., "A generative model for raw audio," 2016, https://arxiv.org/abs/1609.03499.

[16] J. Liu, A. Shahroudy, G. Wang, L. Duan, and A. C. Kot, "SSNet: scale selection network for online 3D action prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 83499–88358, Salt Lake City, UT, USA, June 2018.

[17] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan, and A. Kot Chichung, "NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 6, pp. 1453–1467, 2019.

[18] N. Kalchbrenner, A. van den Oord, K. Simonyan et al., "Video pixel networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, JMLR, Sydney, Australia, pp. 1771–1779, August 2017.

[19] Z. Xu, Y. Wang, M. Long, and J. Wang, "PredCNN: predictive learning with cascade convolutions," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2940–2947, Stockholm, Sweden, July 2018.

[20] X. Liu, J. Yin, H. Liu, and Y. Yin, "PISEP^2: pseudo image sequence evolution based 3D pose prediction," 2019, https://arxiv.org/abs/1909.01818.

[21] V. Bloom, D. Makris, and V. Argyriou, "G3D: a gaming action dataset and real time action recognition evaluation framework," in *Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, June 2012.

[22] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: a large scale dataset for 3D human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.

[23] J. Butepage, M. J. Black, D. Kragic, and A. H. Kjellstrom, "Deep representation learning for human motion prediction and classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6158–6166, Honolulu, HI, USA, July 2017.

[24] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5226–5234, Salt Lake City, UT, USA, June 2018.