LSTformer: Long Short-Term Transformer for Real Time Respiratory Prediction

Min Tan, Huixian Peng, Xiaokun Liang, Yaoqin Xie, Zeyang Xia, *Senior Member, IEEE*, and Jing Xiong, *Member, IEEE*

***Abstract*— Since the tumor moves with the patient's breathing movement in clinical surgery, the real-time prediction of respiratory movement is required to improve the efficacy of radiotherapy. Some RNN-based respiratory management methods have been proposed for this purpose. However, these existing RNN-based methods often suffer from the degradation of generalization performance for a long-term window (such as 600 ms) because of the structural consistency constraints. In this paper, we propose an innovative Long Short-term Transformer (LSTformer) for long-term real-time accurate respiratory prediction. Specifically, a novel Long-term Information Enhancement module (LIE) is proposed to solve the performance degradation under a long window by increasing the long-term memory of latent variables. A lightweight Transformer Encoder (LTE) is proposed to satisfy the real-time requirement via simplifying the architecture and limiting the number of layers. In addition, we propose an application-oriented data augmentation strategy to generalize our LSTformer to practical application scenarios, especially robotic radiotherapy. Extensive experiments on our augmented dataset and publicly available dataset demonstrate the state-of-the-art performance of our method on the premise of satisfying the real-time demand.**

***Index Terms*—Radiation therapy, respiratory prediction, transformer.**

Manuscript received 12 February 2022; revised 7 June 2022 and 13 July 2022; accepted 13 July 2022. Date of publication 18 July 2022; date of current version 5 October 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62073309, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022B1515020042, in part by the Chinese Academy of Sciences Youth Innovation Promotion Association Excellent Member Program under Grant Y201968, in part by the Shenzhen Science and Technology Program under Grant JCYJ20210324115606018, and in part by the Opening Foundation of State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and technology under Grant DMETKF2020016. *(Min Tan and Huixian Peng contribute equally to this work.) (Corresponding authors: Zeyang Xia; Jing Xiong.)*

Min Tan is with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with the University of Chinese Academy of Sciences, Beijing 101400, China (e-mail: min.tan@siat.ac.cn).

Huixian Peng is with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with the School of Mechatronic Engineering and Automation, Foshan University, Foshan 528255, China (e-mail: huixian.peng@qq.com).

Xiaokun Liang, Yaoqin Xie, Zeyang Xia, and Jing Xiong are with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: xk.liang@siat.ac.cn; yq.xie@siat.ac.cn; zy.xia@siat.ac.cn; jing.xiong@siat.ac.cn).

Digital Object Identifier 10.1109/JBHI.2022.3191978

I. Introduction

IN EXTERNAL beam abdominal radiotherapy, respiration is the main factor causing the tumor's movement, which significantly lowers the accuracy of radiotherapy [1]–[4]. The respiratory signal is semi-periodic and nonstationary in clinical practice [5]. Specifically, it changes amplitude and period over time and varies among patients [4]. In addition, radiotherapy apparatuses exhibit inherent system latency [6]–[9] caused by external marker position acquisition, calculation, and mechanical system delay [10]. For instance, the system latency is around 115 ms for CyberKnife [11], and around 500 ms for multi-leaf collimator (MLC) [9], which means that the tumor's movement must be predicted in advance to adjust the radiation beam.

To ensure consecutive radiation and minimize the damage to adjacent tissue, direct tumor tracking methods aim to track the metallic marker implanted inside the tumor [12], [13] or predict the in-plane organ's motion by 2D Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) [4], [14], [15]. By contrast, model-based methods indirectly predict the respiratory signal instead of the 2D image processing for real-time tumor tracking. With higher computational efficiency than direct methods, the indirect respiratory prediction problem has received ever-increasing attention in research and clinical communities. Among them, the Kalman filter (KF) model [16], [17] is widely used because of its good ability to fit time series data. However, KF-based methods need to establish the state equation and observation equation first, while it is difficult to determine these initial parameters [18]. This weakness greatly affects the performance of KF-based methods for nonlinear respiratory signals.

To further improve the model's fitting ability for nonlinear signals, the learning-based method is proposed as an alternative solution, in which the utilization of deep recurrent neural networks (RNN) [19]–[22] to process respiratory signals has become one of the mainstream paradigms. Although there are many improved variants of RNNs, such as the Gated Recurrent Unit (GRU) [22], the issue of how to capture long-term dependencies of intricate respiration remains unresolved. In addition, [23] shows that the generalized feature extraction of RNN is limited by the structural consistency constraints.

Recently, Transformer [24] has been proposed in a non-autoregressive fashion, which can extract more versatile features via the multi-head attention (MHA) and feed-forward network (FFN) with residual structures. Unlike RNN-based methods, Transformer allows encoding any history inputs to improve

the execution of feature extraction. Therefore, the canonical Transformer can capture the irregular patterns with rich feature representation in the respiratory signal. However, due to the space complexity which quadratically grows with the input window length [25], Transformer cannot cope with long-term respiratory prediction in real-time scenarios.

Based on the above observations, we propose an innovative Long Short-term Transformer (LSTformer) for real-time accurate respiratory prediction under a long window. Specifically, a lightweight Transformer Encoder (LTE) is proposed to model the semi-periodicity temporal dependency of respiratory signals with good generalizability, which ensures real-time requirement by lightweight processing. Moreover, a novel Long-term Information Enhancement (LIE) module is proposed to strengthen the long-term memory of the latent variables for long window prediction, in which the latent variables are encoded by the lightweight Transformer. In addition, since the existing public dataset cannot completely cover the actual application scenarios, an application-oriented data augmentation strategy (AOA) is proposed to expand the diversity of the public dataset collected by the optical tracker. We use the depth camera to collect new data on the simulator to train the model, which can improve the model's generalization ability and help the model quickly adapt to real application scenarios, especially robotic radiotherapy.

In summary, our main contributions are three-fold:

- An innovative Long Short-term Transformer is proposed for long-term real-time accurate respiratory prediction. Experimental results verify the superior performance on publicly available datasets and our synthetic dataset by breathing simulator.
- A Long-term Information Enhancement module is proposed to solve the performance degradation problem of RNN-based approaches under a long-term window. In addition, a lightweight encoder structure is proposed to extract features with strong representation capabilities while ensuring real-time requirement.
- An application-oriented data augmentation strategy is proposed to improve the model's generalization performance in different application scenarios via synthesizing datasets created by an artificial breathing simulator.

II. Related Works

A. Direct Tumor Tracking Methodology

Direct tumor tracking methods aim to track the metallic marker implanted inside the tumor or model the in-plane organ's motion using CT or MRI images. Bourque *et al.* [15] proposed a combination of particle filter and autoregressive motion prediction algorithm for real-time MRI-guided lung cancer radiotherapy. Romaguera *et al.* [4] proposed a discriminative spatial transformer network to predict in-plane organ motion. Jafari *et al.* [26] proposed a lung biomechanical model for predicting tumor motion using lung 4D CT scans of radiation treatment planning. However, the continuous imaging of internal tumor's motion is challenging due to system latencies [4]. Besides, the 2D image sequences processing can be time-consuming, especially for high resolution and large datasets [4]. These reasons weaken the performance of direct tumor tracking algorithms in real-time systems.

B. Indirect Model-Based Respiratory Prediction Methodology

Respiratory prediction belongs to the time series forecasting problem. Traditional statistical forecasting methods predict the future series data on the basis of historical data [27]. McCall *et al.* [28] proposed an autoregressive moving average (ARIMA) algorithm to define a mathematical model of respiratory motion's periodic and non-periodic components. This model assumes that the data are linear and follow a specific probability distribution. Putra *et al.* [16] proposed an interacting multiple model (IMM) filter and two KF models. Vedam *et al.* [29] evaluated the predictive power of sinusoidal and adaptive filters by calculating the standard deviation of the instantaneous differences between the predicted and actual breathing positions. Ruan *et al.* [5] used the semi-periodicity characteristic of respiratory motion to train the local regression model from previous observation data. Wu *et al.* [30] utilized finite-state model to describe three respiratory states. Çetinkaya *et al.* [31] used the exact KF to model the state estimation of quasi-periodic respiratory signals. Hong *et al.* [17] proposed a cascade structure of an extended KF and support vector regression (SVR), and this cascade structure proved to be more effective than SVR and artificial neural network (ANN) alone. Ernst *et al.* [32] evaluated several respiratory prediction algorithms such as the extended KF, wavelet-based multi-scale autoregression (wLMS), and ϵ-support vector regression (SVRpred) methods. They found that the wLMS is the most effective algorithm. Bao *et al.* [33] proposed a patient-specific respiratory motion model based on the bayesian and PCA algorithm. Jöhl *et al.* [34] compared 18 prediction filters to evaluate the sufficiency of linear filters. Model-based methods are computationally efficient and can be responsive to irregular changes. However, they often suffer from poor performance for signals since the respiratory motion is nonlinear, semi-periodic, and nonstationary in clinical practice [5].

C. Indirect Learning-Based Respiratory Prediction Methodology

Learning-based methods with high nonlinearity are proposed to improve the prediction accuracy of respiratory motion. Seregni *et al.* [35] proposed an ANN model for phantom testing to demonstrate the feasibility of real-time tumor tracking based on external respiratory signals. Sun *et al.* [36] applied adaptive boosting and multi-layer perceptron neural network (ADMLP-NN) to improve the prediction accuracy of respiratory signals. Teo *et al.* [37] focused on online neural network optimization and concluded a three multi-layer perceptron (MLP) neural network. Subsequently, Sun *et al.* [38] investigated the prediction performance of four popular adaptive boosting methods based on the MLP-NN. Chang *et al.* [39] developed a temporal convolutional neural network (TCN) to predict respiratory signals and measured the RMSE in three-dimensional space. Because ANN ignores the temporal dependency, RNN-based methods have emerged. Lin *et al.* [21] firstly introduced long short-term

TABLE I
DATA DETAILS

Dataset	Object	Collect Device	Data Amount [a]
Public	38 Patients	Optical Tracker	7500×100
Augmentation	Phantom Device	RGB-D Camera	7500×6

[a]This data term represents the dimensions of the respiratory signal matrix, where 100 and 6 refer to the number of respiratory signals. 7500 refers to the amplitude of motion value, corresponding to 7500 timestamps.

memory (LSTM) [40] to solve the respiratory prediction problem in 500 ms. Wang *et al.* [41] also proposed to use LSTM to predict external respiratory motion signals and then built the external/internal motion correlation models. LSTM is developed to selectively remember and update important information by special forgetting and saving mechanism. Wang *et al.* [20] proposed bidirectional LSTM (Bi-LSTM) to further improve tracking accuracy, where the predictive window is in the range of 40 ms to 400 ms. Yu *et al.* [22] proposed a Bi-GRU, another variant of RNN, to mitigate the long updating time problem of LSTM.

Because of the limitations such as the structural consistency constraints [23] and weak long-term dependencies [25], RNN-based methods are not effective in respiratory prediction under long windows. Although Transformer [24] can extract generated and robust features, it cannot be directly used in real-time respiratory prediction because of the high space complexity. Thus, we simplify the Transformer Encoder and propose a novel LIE module to realize real-time prediction of respiratory signals.

III. MATERIALS AND METHODS

A. Application-Oriented Data Augmentation

Since the collection of respiratory data is affected by the collection equipment and environment, different collection equipment (such as optical tracker [11], real-time position management (RPM) system [42] and RGB-D camera [43].) will show a slight difference on the same person. In order to make our model have good practical application value, we propose an application-oriented data augmentation strategy (AOA) to expand the diversity of public dataset collected by the optical tracker. Under AOA our LSTformer can be generalized to practical application scenarios, especially robotic radiotherapy. The training dataset consists of the public dataset and our augmented dataset, as shown in Table I.

Public Dataset: The public dataset[1] is collected by Dr Kevin Cleary and Dr Sonja Dieterich during CyberKnife treatment at Georgetown University Hospital [44]. The origin database consists of 304 motion traces of 38 patients' respiratory motion between 6.5 and 132 minutes with an acquisition frequency of 26Hz [45].

Augmented Dataset: To increase the diversity of the public dataset and improve the generalizability of the trained model, we obtained the augmented dataset via a breathing simulator.

Finally, we trained our LSTformer model by randomly picking 97 training data segments (90% dataset). The rest of the 9 data segments (10% dataset) are for testing, and each segment contains 7500 points of respiratory motion.

Algorithm 1: Data Augmentation.

Input: Depth image $D \in \mathbb{R}^{H\times W}$, RGB imgae $I \in \mathbb{R}^{H\times W}$, k, ε, n, threshold
Output: $depth$

$N = \{0\}, n_i \in N, i = \{1, 2, \ldots, k\}$
$W = \{w_1, w_2, \ldots, w_k\}$
Iteratively get pixel p from D with the value of NaN and $neighbors$ of p
if $distance(p, neighbor) < \varepsilon$ **then**
  $n_i \leftarrow Select(neighbor)$
end if
Update $p \leftarrow \frac{n_1w_1+n_2w_2+\cdots\cdots+n_kw_k}{w_1+w_2+\cdots\cdots+w_k}$
Iteratively get pixel p from I
if $gray(p) >$ threshold**then**
  Get binary imgae I_B
end if
Use morphological opening to remove isolated dots and burrs
$I_O \leftarrow Open(I_B)$
Figure out the center point coordinates of each marker M
$M = GetConnectedRegion(I_O)$
for i in $1, \ldots, n$**do**
  Center coordinate $(x_i, y_i) \leftarrow Caculate(M_i)$
  $depth_i \leftarrow D(x_i, y_i)$
end for

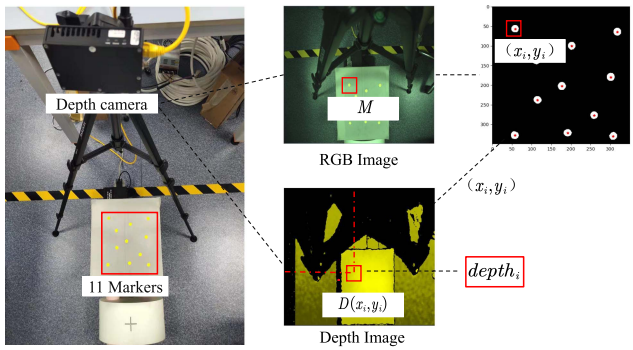

Fig. 1. Breathing simulator data collecting and processing. The (x_i, y_i) denotes the center pixel coordinate of marker M in the RGB image. The $D(x_i, y_i)$ is the coordinate in the depth image and the $depth_i$ is the depth value.

1) Augmentation Data Collection: We made an augmentation dataset utilizing an RGB-D camera to collect motion signals in a breathing simulator phantom device.[2] It is worth noticing that the movement of the simulator is driven by the clinical patient's respiration. Other details can be seen in our previous work [46].

Fig. 1 illustrates the process of breathing simulator data collecting. The 11 markers are placed in the abdominal phantom.

[1][Online]. Available: https://signals.rob.uni-luebeck.de/index.php

[2]A general physical radiotherapy simulator for CIRS Tissue Simulation & Phantom Technology https://www.cirsinc.com/

The RGB-D camera is placed nearly 73 cm above the breathing simulator to collect the abdomen area's depth image and corresponding RGB image. Firstly, the K-nearest neighbors method [47] is applied to fill the gaps and missing values in the depth image. Then the global binary threshold algorithm (Otsu's method) [48] is used to locate all 11 marker points on the RGB image. After that, we perform contour detection and ellipse fitting steps to accurately identify the center coordinates of each marker to obtain depth value. The specific processing flow is shown in Algorithm 1. For breathing simulator data, 20 collective experiments for 11 markers were performed to record the respiratory amplitude. For each experiment, we randomly chose one marker's motion data as one sample, including about 2000 to 3000 points. The final augmented dataset will be reshaped to a size of 7500×6, as shown in Table I, and mixed into the public dataset to improve the generalizability of our model to deal with actual complex scenarios.

2) Data Preprocessing: To alleviate the influence of the acquisition environment and random noise in the sensor, we performed data preprocessing on the premise of maintaining the original signal shape and period characteristics. The data preprocessing includes denoising, smoothing, normalization and partitioning.

Denoising: There are some outlier points and sharp peaks in the respiratory signals. We used the Boxplot method [49] to remove these significant outliers $D_{outliers}$. Boxplot calculates the quartiles of the data and divides all the data into four quartiles in ascending order, which can be formulated as:

$$D_{outliers} = \{x \mid x \notin [Q_1 - \mu(Q_3 - Q_1), Q_3 - \mu(Q_3 - Q_1)]\} \tag{1}$$

where Q_1 and Q_3 are the lower and upper quartiles calculated by Boxplot. μ is the scale parameter which is set to 1.0 from experience in our experiment.

Smoothing: To facilitate our network's processing of discrete respiratory movement data, we employ Savitzky-Golay (SG) filter [50] to smooth the data and filter some white noise. SG filter is a convolution fitting algorithm based on the least-squares method, which can retain the shape and width of the respiratory signal after filtering. The j^{th} smoothed data point Y_j can be formulated as:

$$Y_j = \sum_{i=\frac{1-m}{2}}^{\frac{m-1}{2}} C_i y_{j+i}, \quad \frac{m+1}{2} \le j \le n - \frac{m-1}{2} \tag{2}$$

where y is original discrete respiratory data, Y is smoothed respiratory data, m and n are the smoothing window size and signal length, respectively, and C is the convolution coefficient.

Normalization: To ensure the stability of the model, we normalize the one-dimensional data into the range of 0 to 1. During the testing process, a de-normalized operation has been done to compute evaluative criteria.

Partitioning: Based on the demand of the respiratory motion prediction task, we divide the collected data via our partitioning strategy. Specially, we divide each respiratory sample into many pieces, including input data of length n and target data of length N_{output} in a fixed-size window. The details are shown in Fig. 2.

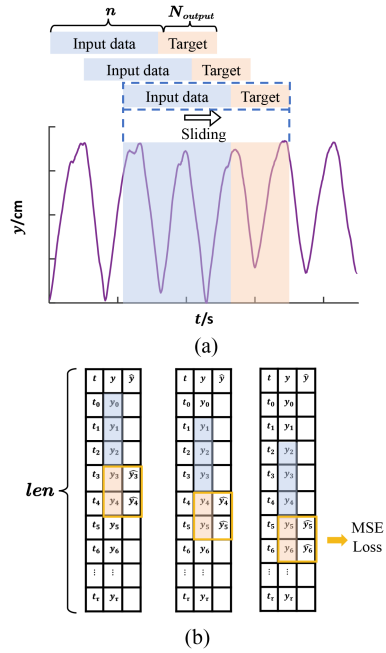

Fig. 2. The partitioning process of respiratory data. (a) In the sliding window with a fixed size (marked with a blue dotted box), n data points are input into the model to predict these N_{output} points. (b) A visual illustration of the partitioning process for $n = 3$ and $N_{output} = 2$, where the len is the length of the time axis, y and $\hat{y}$ are the ground truth data and predictive data, respectively. The MSE loss is calculated to optimize these weights in our model.

B. Long Short-Term Transformer

To further improve the feature extraction capability in respiratory motion prediction and the performance under a long window, we propose an innovative LSTformer for long-term real-time accurate respiratory prediction (as shown in Fig. 3). Specifically, a lightweight Transformer Encoder is proposed to extract generalized feature representations while ensuring real-time requirement. Moreover, a novel Long-term Information Enhancement module is proposed to solve the problem of performance decline under a long window by increasing the long-term memory of latent variables.

1) Lightweight Transformer Encoder: The Transformer [24] is utilized as a powerful autoencoder to tackle the text translation tasks, which consists of three core modules, namely the multi-head attention module (MHA), the feed-forward network (FFN), and the Positional embedding (PE). To retain the feature extraction capability of the Transformer while meeting the real-time requirement, we propose a lightweight Transformer Encoder (LTE, shown in Fig. 4) via adjusting these three core modules specifically for respiratory data.

Since the respiration curve is a one-dimensional feature of time, we define an embedding layer to expand the shape of the respiration features. Specifically, for every input vector $x = [x_0\ x_1\ x_2\ \dots\ x_n]$, we expand the shape of x to $n \times d$ dimensions by the input embedding, as shown in Fig. 4. The embedding layer can be formulated as:

$$x' = L_{ie}(x; W_{ie}), \{x' \in \mathbb{R}^n \times \mathbb{R}^d, x \in \mathbb{R}^n\} \tag{3}$$

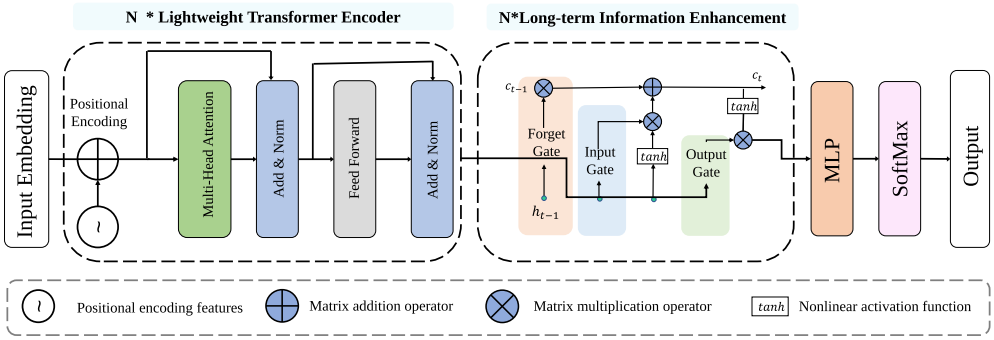

Fig. 3. The architecture of our Long Short-term Transformer (LSTformer). The N-layer asymmetric autoencoder pipeline as the backbone predicts the breathing curve under the long window using historical respiratory data. In Lightweight Transformer Encoder, respiratory data needs to be converted into the nonlinear features via Input Embedding and Positional Encoding (detailed in Fig. 4(a)). Long-term Information Enhancement module mainly includes the Forget Gate, Input Gate, and Output Gate (detailed in (8)-10), where c denotes the cell state during temporal sequence transmission, h denotes the hidden state, and $tanh$ denotes the activation function.

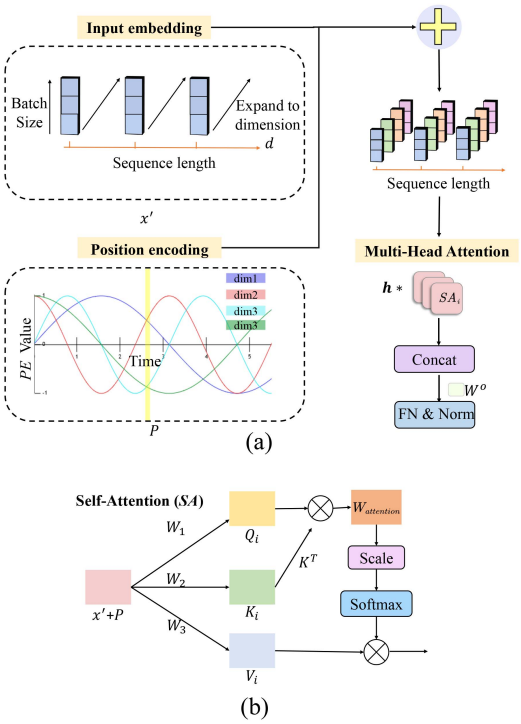

Fig. 4. The architecture of lightweight Transformer Encoder (LTE) and Self-Attention. (a) The logical relationship between sub-modules in LTE. Specifically, x' represents the expanded respiratory data by Input Embedding, P denotes the learnable position features encoded by segmentation embedding, h and W^O are the header number and the header weights of MHA blocks, respectively. The operations of Concat, FN and Norm are shown in eq.6 and eq.7. (b) Calculation of the Self-Attention (SA), where the $W_{\{1,2,3\}}$ is the learnable weight, the queries (Q_i), keys (K_i), and values (V_i) are transformations of the corresponding input state vectors to calculate the weight matrix $W_{attention}$ (detailed in eq.5). The final output of SA is obtained by Scale and Softmax operations.

where x' denotes the expanded inputs, L_{ie} denotes the input embedding layer, W_{ie} denotes the embedding weight matrix, n denotes the dimension of origin input x, and d denotes the expanded embedding dimension.

Since the respiration data is semi-periodic, using the original PE to add position coding for each point on the respiration curve will bring redundancy and increase the computational cost. Hence, we propose a learnable segmentation embedding to add positional information to x' adaptively, where a learnable mask will determine which segments would be reserved for PE operation. The learnable segmentation embedding can be formulated as:

$$P(i,j) = \begin{cases} \frac{1}{S}\sum^{S} \sin\left(\frac{i}{10000^{j/d}}\right), & \text{if } j = 2\,k \\ \frac{1}{S}\sum^{S} \cos\left(\frac{i}{10000^{j/d}}\right), & \text{if } j = 2k+1 \end{cases} \tag{4}$$

where S represents the learnable segment spacing, i and j represent the position in the time axis and the current dimension of embedding, respectively. $i \in [0, n]$ and $j \in [0, d]$. The value of the matrix P computed by sine or cosine function varies depending on whether the subscript j is odd or even, $k \in \boldsymbol{Z}$.

Multi-head attention is made up of h self-attention (SA, shown in Fig. 4(b)) blocks. Given the positional embedding matrix of $x' + P$ as the input for each head SA_i, the calculation of the SA mechanism requires three matrices Q_i (query), K_i (key) and V_i (value), which are obtained by applying three different linear transformations for $x' + P$. Considering the real-time requirement, we simplified the implementation of original SA via leveraging single-layer MLP to realize the linear transformation. The SA mechanism can be formulated as:

$$SA_i = Softmax\left(Q_i(x' + P; W_i^Q)K_i(x' + P; W_i^K)^T\right)\frac{V_i(x' + P; W_i^V)}{\eta} \tag{5}$$

where the $W_i^Q \in \mathbb{R}^{d\times d_k}$, $W_i^K \in \mathbb{R}^{d\times d_k}$, and $W_i^V \in \mathbb{R}^{d\times d_k}$ denote the linear transformation weights, and η is the scaling factor.

MHA learns rich features without increasing the time complexity by integrating multiple SA_i. The final MHA feature can be formulated as:

$$\mathrm{MHA} = \mathrm{Concat}\left(SA_1, SA_2, \ldots, SA_h\right) W^O \tag{6}$$

where the $W^O \in \mathbb{R}^{d\times hd_k}$ is the header weights.

After the multi-head attention, the position-wise fully connected with feed-forward network (FN) and normalization (Norm) are followed closely, as shown in Fig. 4. The calculation

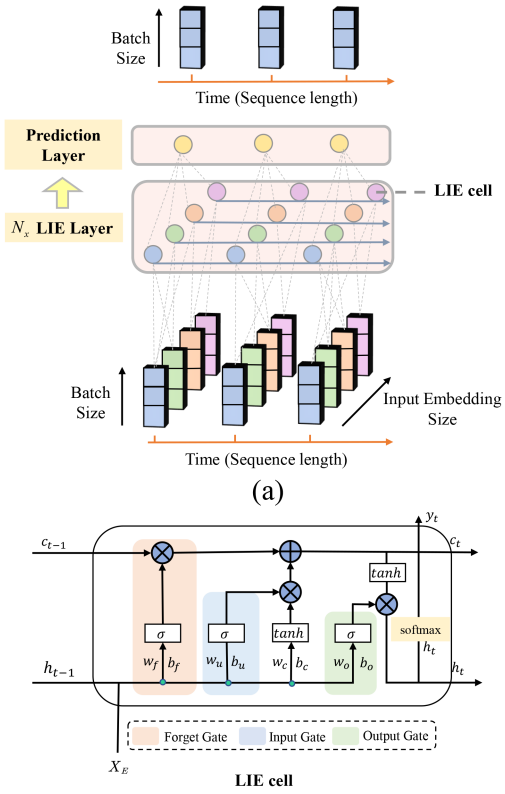

Fig. 5. The architecture of Long-term Information Enhancement (LIE). (a) The decoding process of LIE, where the data cuboids at the bottom denote the outputs from the LTE, N_x denotes the number of LIE layers, and the prediction layer is an MLP. (b) The calculation process of each LIE cell, where $c_{\{t-1,t\}}$ denotes the cell state in one time step, $h_{\{t-1,t\}}$ denotes the hidden states, and X_E denotes the input data. $w_{\{f,u,b\}}$ and $b_{\{f,u,b\}}$ denotes the weights and biases corresponding to different Gates, respectively (see eq.8-10 for details).

of the LTE output can be formulated as:

$$X_E = \text{Norm}(W_2\text{ReLU}(W_1\text{MHA} + b_1) + b_2) \tag{7}$$

where W_1 and W_2 represent transformation weights, b_1 and b_2 represent transformation biases, and $\text{ReLu}(\cdot)$ is the nonlinear activation function.

2) Long-Term Information Enhancement: Although the Transformer can obtain the latent variables with rich feature representation, it cannot establish effective long-term effects [25]. To improve the accuracy of respiratory prediction under the long window, we combine LSTM cells and MLP, and propose a Long-term Information Enhancement (LIE, shown in Fig. 5(a)), which aims to strengthen the long-term memory of latent variables and predict respiratory signals. As shown in Fig. 5(b), the LTE cell consists of three gates to control updating cell state c_t and hidden state h_t in the LIE layer. Among them, the input gate controls how much of the current input X_E is saved to the cell state c_t, the forget gate determines how much of c_{t-1} should be remembered in current moment for c_t, and the output gate outputs features enhanced by long-term information, which can be formulated as:

$$G_t^{input} = \sigma(\text{Concat}(h_{t-1}, X_E); W) \times \tau(\text{Concat}(h_{t-1}, X_E); W) \tag{8}$$

$$G_t^{forget} = \sigma(\text{Concat}(h_{t-1}, X_E); W) \tag{9}$$

$$G_t^{output} = h_t = \sigma(\text{Concat}(h_{t-1}, X_E); W) \times \tau(c_t) \tag{10}$$

where σ is the sigmoid function, τ is the $\tan(\cdot)$ activation function, and W is the weight of corresponding functions. The final cell state c_t satisfies the equation $c_t = G_t^{forget} \times c_{t-1} + G_t^{input}$.

We only keep the last moment output value y_t in the output layer and add the MLP layer composed of linear transformation nodes. Finally, the softmax layer converts the output of the decoded block into a probability as the final output. The final decoded results can be formulated as:

$$\mathcal{Y} = \frac{e^{(W_D \cdot y_t + b_D)}}{\sum_{i=1}^{C} e^{(W_D \cdot y_t + b_D)}} \tag{11}$$

where W_D and b_D are the weight and bias of the MLP layer, respectively. The last moment output is $y_t = Softmax(h_t; W)$.

IV. Experiment

A. Implementation Details

We implement our LSTformer to predict the long-term respiratory curve from the one-dimensional respiratory sequence via LTE and LIE. In LTE, we utilize a 64-dimensional embedding layer for the input embedding, four headers for the multi-head attention, the MLP with 2048, and 64 hidden units for the feed-forward. There are two LSTM core structures for the long-term enhancement cell and one MLP with 64 and 1 hidden unit for the prediction layer in LIE. Our approach is implemented with Pytorch 1.5.0 and trained for 200 epochs by the Adam optimizer [51]. For all datasets, the batch size is set to 64, and the learning rate is set to 0.5e-2 for training LSTformer. All experiments in this paper were performed on NVIDIA RTX 2080 GPU with 8 G memory and Intel Xeon E5 CPU with 2.30 GHz. During the training process, mean square error (MSE) loss and backwards gradients are calculated to optimize the weight parameters, which can be formulated as:

$$MSE_{loss}(\hat{y}, truth) = \frac{1}{N_{output}} \sum (\hat{y}_i - truth_i)^2 \tag{12}$$

where $\hat{y}$ is the predicted respiratory curve by our method, $truth$ is the ground-truth respiratory curve, and N_{output} is the length of predicted points (as shown in Fig. 2).

B. Evaluation Protocol

For a fair comparison, we adopt three evaluation metrics in the experiment. The details are listed as follows:

- $MAE(mm)$: The mean absolute error between predictive curve and truth, which can be formulated as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \|y_i - \hat{y}_i\| \tag{13}$$

- $RMSE(mm)$: The root mean squared error between predictive curve and truth, which can be formulated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{14}$$

- $SIM_{dtw}(mm)$: The similarity between predictive curve and truth in shape. This metric exploits dynamic time warping algorithm (DTW) [52] to calculate the least cost $Path$, which can be formulated as:

$$SIM_{dtw}(X,Y) = \sqrt{\sum_{(i,j)\epsilon Path}(X_i - Y_j)^2} \tag{15}$$

where i and j represent the subscripts of the two-time series X and Y, respectively. The final DTW similarity value SIM_{dtw} is obtained by summing the Euclidean distances of the corresponding points in the minimum cost path.

TABLE II
DETAIL EVALUATION CRITERIA VALUES OF LSTFORMER MODEL IN DIFFERENT TIME LATENCY

latency /Model	$ave.SIM_{dtw}\downarrow$ (mm)	$ave.MAE\downarrow$ (mm)	$ave.RMSE\downarrow$ (mm)	$ave.Time\downarrow$ (ms)
200(N_{output} =5)				
Lin *et al.* [21]	22.445	0.325	0.421	**14.53**
Bi-LSTM [20]	17.698	0.425	0.515	27.09
Bi-GRU [22]	22.770	0.399	0.558	28.52
Ours	**11.932**	**0.241**	**0.323**	22.29
400(N_{output} =10)				
Lin *et al.* [21]	28.050	0.427	0.527	**29.53**
Bi-LSTM [20]	48.258	1.003	1.154	57.49
Bi-GRU [22]	44.503	0.644	0.808	56.82
Ours	**17.774**	**0.344**	**0.454**	44.19
600(N_{output} =15)				
Lin *et al.* [21]	91.326	1.070	1.314	**44.36**
Bi-LSTM [20]	62.543	1.254	1.456	86.59
Bi-GRU [22]	123.132	1.346	1.577	84.95
Ours	**18.503**	**0.355**	**0.501**	66.09

↓ Indicates That the Smaller the Value, the Better.

C. Comparing With State-of-The-Art Methods

To verify the effectiveness of the proposed LSTformer in respiratory prediction, we compared our model with three state-of-the-art (SOTA) models. For these SOTA methods [20]–[22], we replicate them following the same configuration. For instance, in Lin's method [21] and Bi-GRU [22], the number of layers is 3, and in Bi-LSTM [20], the number of layers is 7. In addition, the learning rate and optimizer also keep the same configuration to ensure optimal performance.

The experiments are executed in situations of different predictive points (N_{output} =5, 10, 15) to prove the superiority of our method under the long window. The different output number of points corresponds to different time latency. For instance, the respiratory motion points after 200 ms latency can be predicted when $N_{output} = 5$, and those after 600 ms latency can be calculated when $N_{output} = 15$, respectively. It can be formulated as follows:

$$latency = 1000\frac{1}{f} \tag{16}$$

where the $f = 26$ Hz is the signal acquisition frequency. We multiply 1000 to obtain the latency time with the unit of millisecond.

The detailed results of different latency windows are listed in Table II. In our method, the SIM_{dtw} can be reduced to 11.932 mm, 17.774 mm and 18.503 mm in the latency window of 200 ms, 400 ms and 600 ms, respectively. For the latency window of 200 ms, our method outperforms the benchmark of Lin *et al.* [21] by 10.513 mm on SIM_{dtw}, by 0.084 mm on MAE, and by 0.098 mm on $RMSE$. Our method outperforms Bi-LSTM by 5.766 mm on SIM_{dtw}, by 0.184 mm on MAE, and by 0.192 mm on $RMSE$. Our method outperforms Bi-GRU by 10.838 mm on SIM_{dtw}, by 0.158 mm on MAE, and by 0.235 mm on $RMSE$. For the latency window of 400 ms, our method outperforms the benchmark of Lin *et al.* [21] by 10.276 mm on SIM_{dtw}, by 0.083 mm on MAE, and by 0.073 mm on $RMSE$. Our method outperforms Bi-LSTM by 30.484 mm on SIM_{dtw}, by 0.669 mm on MAE, and by 0.7 mm on $RMSE$. Our method outperforms Bi-GRU by 26.729 mm on SIM_{dtw}, by 0.3 mm on MAE, and by 0.354 mm on $RMSE$. Especially in the longest latency window of 600 ms, our method outperforms the benchmark of Lin *et al.* [21] by 72.823 mm on SIM_{dtw}, by 0.715 mm on MAE, and by 0.813 mm on $RMSE$. Our method outperforms Bi-LSTM by 44.04 mm on SIM_{dtw}, by 0.899 mm on MAE, and by 0.955 mm on $RMSE$. Our method outperforms Bi-GRU by 104.629 mm on SIM_{dtw}, by 0.991 mm on MAE, and by 1.076 mm on $RMSE$. Besides, the last column of Table II shows the evaluation of the real-time requirement. It can be observed that all models can predict results within 100 ms under the latency window of 600 ms. However, our method significantly outperforms Bi-LSTM and Bi-GRU in multi-scale prediction accuracy while ensuring lower time-consuming. Although our method slightly underperforms Lin's method (by 6.2 ms on 200 ms latency, by 13.3 ms on 400 ms latency, and by 19.5 ms on 600 ms latency), our method achieves great performance on prediction accuracy under multi-scale latency windows. In summary, our method achieves a preferable performance and outperforms other RNN-based methods under long window prediction while satisfying the real-time requirement.

D. Ablation Study

In this section, we first conduct experiments to verify the effectiveness of our proposed components, including the backbone, LTE, LIE and PE. We further analyze the impact of the model's structure on the performance. Then, we investigate the effect of augmented data for discussion. Finally, we demonstrate the stability of our LSTformer under long-term window prediction.

Effectiveness of PE, LTE and LIE: We present the results under different combinations of our proposed modules in LSTformer for long-term window prediction, including Trans-Encoder (only LTE), original Transformer (Trans-Original), Pure-LSTM (only LIE), and our LSTformer. The detailed information is listed in Table III. From Table III, we can observe that our LSTformer outperforms Pure-LSTM by 72.823 mm on $ave.SIM_{dtw}$ criteria, by 0.372 mm on MAE and by 0.354 mm on $RMSE$.

TABLE III
ABLATION ON DIFFERENT MODULE COMBINATIONS

Model	LTE	LIE	PE	$SIM_{dtw}\downarrow$ (mm)	$MAE\downarrow$ (mm)	$RMSE\downarrow$ (mm)	$Time\downarrow$ (ms)
Trans-Original	✗	✗	✗	43.467	1.008±0.552	1.182±0.658	214.9
Trans-Encoder	✓	✗	✓	41.881	0.727±0.347	0.855±0.416	56.54
Pure-LSTM	✗	✓	✗	91.326	1.070±0.538	1.314±0.667	**44.36**
LSTformer-no-PE	✓	✓	✗	175.579	1.817±0.841	2.102±0.981	54.33
LSTformer	✓	✓	✓	**18.503**	**0.355±0.218**	**0.501±0.270**	66.09

The Results are Arithmetic Mean on 9 Test Segments Under 600 Ms Latency Window.

Compared with Trans-Encoder, LSTformer outperforms it by 23.378 mm on $ave.SIM_{dtw}$ index, by 0.715 mm on MAE and by 0.813 mm on $RMSE$. Compared with Trans-Original, LSTformer outperforms it by 24.964 mm on $ave.SIM_{dtw}$ index, by 0.653 mm on MAE and by 0.681 mm on $RMSE$. Since the original Transformer is a complete encoder-decoder architecture and requires additional mask operations to cover the information predicted in the testing, the computation time increases significantly (214 ms). By contrast, our LTE and LIE can optimize the time efficiency, which improves by 69.24%. Compared with LSTformer-no-PE, LSTformer outperforms it by 157.076 mm on $ave.SIM_{dtw}$ index, by 1.462 mm on MAE and by 1.601 mm on $RMSE$. It means that our PE can improve prediction accuracy. In summary, our proposed modules can significantly improve the performance of long-term window prediction compared with Pure-LSTM and original Transformer.

Analysis of Structure: To analyze the effect of different architecture, we test the accuracy of LTE and LIE modules under different numbers of layers, encoding dimensions, and MHA heads. The detailed information is listed in Fig. 6. Our model achieves the best performance from the experimental results under the combination of two-layer LTE, two-layer LIE, 64-dimensional encoding, and 4-heads MHA. It means that our LSTformer is lightweight while ensuring optimal performance.

Generalizability of Augmentation Data: Fig. 7 demonstrates that data augmentation brings great generalization and predictive ability to LSTformer when facing different collection scenarios. LSTformer provides feasibility and reliability for future experiments on the simulator phantom and practical clinical application.

Model Stability under Long-term Window: To demonstrate the stability of our LSTformer under long-term window predictions, we compare the prediction of different methods at 600 ms latency (as shown in Fig. 8). The box plots demonstrate that our LSTformer has outstanding performance on SIM, MAE and $RMSE$. The values of the standard deviation of LSTformer in three indicators are 9.835 mm, 0.217 mm, and 0.270 mm, respectively. To be concrete, LSTformer outperforms Lin *et al.*'s method by 72.82 mm, Bi-LSTM by 44.04 mm, and Bi-GRU by 104.63 mm on SIM_{dtw}. Our method outperforms Lin *et al.*'s method by 0.777 mm, Bi-LSTM by 1.002 mm, and Bi-GRU by 1.037 mm on MAE indicators. Our method outperforms Lin *et al.*'s method by 0.785 mm, Bi-LSTM by 1.010 mm, and Bi-GRU by 1.056 mm on $RMSE$ indicators. It can be concluded that our method can provide more stability than other RNN-based methods under long window prediction.

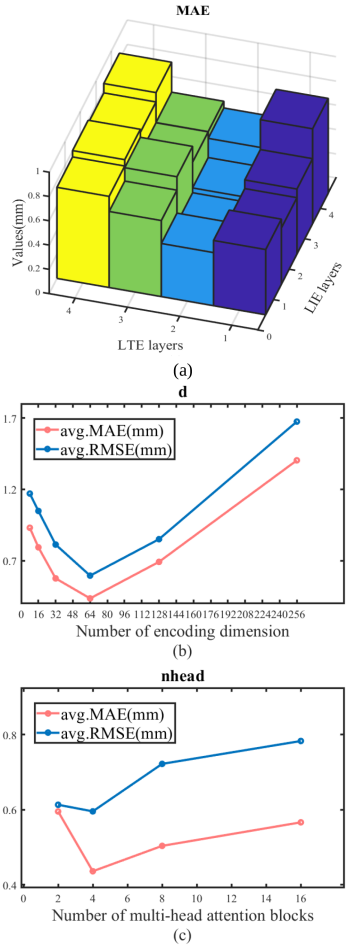

Fig. 6. Inference about hyper-parameters in LSTformer. (a) Inference on numbers of LTE layer and LIE layer. (b) Inference on encoding dimension d. (c) Inference on layer numbers multi-head attention blocks $nhead$.

E. Visualization

As shown in Fig. 9, LSTformer with attention mechanism is used to extract feature representation so that the performance of our model is better than RNN methods in all windows. In addition, the experimental results under the 600 ms window also prove that the LIE module authentically has the ability of long-term information enhancement. In contrast, the feature extraction ability of LSTM is not exceptional enough, so it is incapable of learning the amplitude information of breathing movement well. It can be observed that the Bi-LSTM and Bi-GRU have phase drift phenomena on the time axis, which becomes more evident as the window becomes large. In addition, the two bidirectional models have a poorer shape fitting capacity than Lin *et al.*'s method when irregular spikes appear in the respiratory signal, as shown in the first and third row of Fig. 9. Thus, they often lead to spiculate shapes in the predicted results. Bidirectional models additionally train a model from the future to the past and theoretically concerns more future information in context. However, their complicated structure has generated overfitting phenomenon compared to the mono-directional LSTM approach from the testing results on periodic breathing signals.

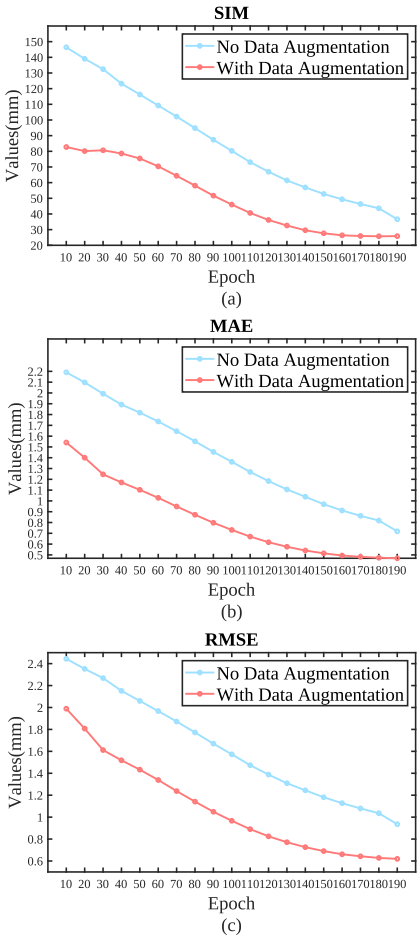

Fig. 7. Comparison of three indicators before and after data enhancement. (a) SIM criteria comparison changes in training process of 190 epochs. (b) MAE criteria comparison. (c) RMSE criteria comparison.

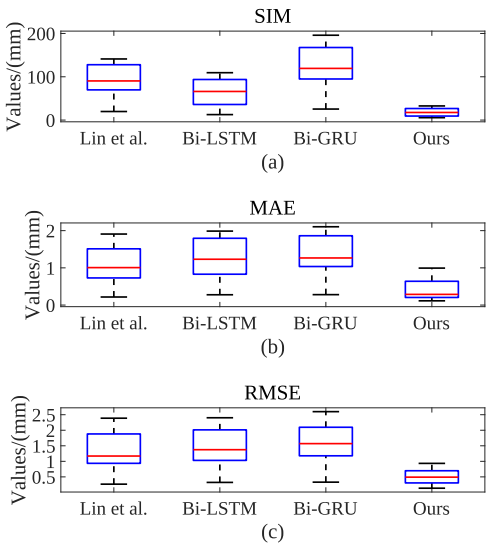

Fig. 8. Stability analysis under long-term window. (a) SIM criteria comparison among RNN-based methods and ours. (b) MAE criteria comparison. (c) RMSE criteria comparison.

V. Discussion

The purpose of this work is to establish a model that can predict the semi-periodic respiratory signals under long-term latency while ensuring real-time requirement. Based on our model, the tumor's motion can be correlated and tracked in radiotherapy. The extensive experiments on three different latency windows demonstrate that our proposed method obtains great prediction accuracy under multi-scale latency, specifically for long-term respiratory prediction. Compared with RNN-based predictive methods, our method achieves a state of the art performance while ensuring the real-time requirement. This paper concerns the respiratory prediction of external abdomen markers, since the processing of internal tumor motion through direct imaging by MRI and CT can be time-consuming [4]. Also, the registration needs to be considered between consecutive images [4]. Therefore, this indirect surface markers predictive method brings advantages of noninvasive and high computational efficiency.

A preoperative correlation model is generated prior to the respiratory prediction in clinical flows. The process is as follows: (1) External breathing motion is monitored by an optical tracker [11], real-time position management system (RPM) [42] or RGB-D camera [43]. (2) Internal tumor motion is measured at multiple discrete time points by acquiring intraoperative X-ray, MRI or CBCT images [53]. (3) A correlation model, described as a simple linear or quadratic function, is generated by fitting the 3D internal tumor motion for different respiratory cycles to the external marker positions. (4) Taking real-time respiratory data as input, the respiratory prediction model quickly outputs predictive data associated with internal tumor position, thus compensating for system delays.

The respiratory signals obtained via a respiratory sensor could detect the periodic motion of a tumor accurately [54] since external respiration has a strong correlation with tumor motion [7], [55]. Therefore, the correlation model between the external marker and internal tumor will be addressed in future studies. Nowadays, several works have investigated the correlation model, such as [41], [43]. However, there are still some issues that have not been completely solved. For example, since the tumor motion is three-dimensional, the influence of respiratory motion is different in three directions. Moreover, other factors, such as baseline shifts of respiration [56] and deformation of tumor [4], [57], also need to be further considered in future research.

One point that deserves discussion is the accuracy level of the radiotherapy application. Our prediction model got an MAE of 0.241 mm in 200 ms latency, 0.344 mm in 400 ms, and 0.355 mm in 600 ms. This result demonstrates the stability of our method in prediction accuracy. The CyberKnife [11] is robotic equipment for clinical radiotherapy, which represents the advanced level. It exploits a stereo camera and orthogonal X-ray to acquire external respiratory signals and internal tumor motion images. CyberKnife can achieve a system error of approximately 0.3 mm to 1.5 mm in latency of around 115 ms to 192.5 ms [7], [53]. Therefore, it could be concluded that our method can obtain a relatively lower error, which may have potential applicability.

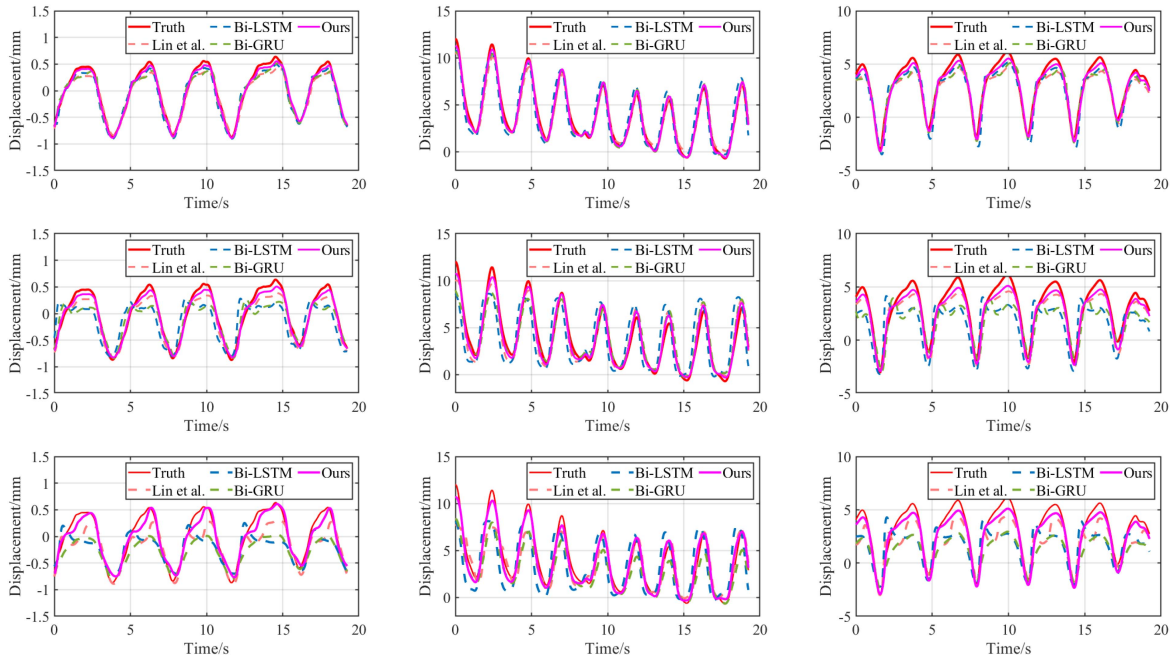

Fig. 9. Predicted trace and ground truth. 1st to 3 rd row: the latency of 200 ms, 400 ms, and 600 ms, respectively. Data source: The first column of the test data above is from timestamp 5721-6201 of test data segment No.1, and the displacement range is from −1 mm to 0.8 mm. The second column is from timestamp 5721-6201 of test data segment No.3, and the displacement ranges is from −1 mm to 12 mm. The third column is from timestamp 6499-6979 of test data No.4, and the displacement range is from −5 mm to 7 mm.

VI. Conclusion

In this paper, an innovative LSTformer is proposed for long-term real-time accurate respiratory prediction. It employs the lightweight Transformer to capture the global temporal semi-periodicity feature of respiratory signals, which is superior to RNN-based methods. Moreover, a LIE module is developed to solve the problem of generalization decline under a long window. An application-oriented data augmentation strategy is used to generalize our LSTformer to practical application scenarios. Extensive experiments on augmentated datasets and publicly available datasets demonstrate the state-of-the-art performance of our method and verify that the attention mechanism of the Transformer can effectively improve the prediction accuracy on the premise of satisfying the real-time demand. The proposed method is general and could be extended to tumor tracking in different clinical scenarios. Future clinical implementation of the proposed method with thorough assessment may provide the physicians with an important tool to assist patient treatment delivery.

References

[1] E. A. Barnes et al., "Dosimetric evaluation of lung tumor immobilization using breath hold at deep inspiration," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 50, no. 4, pp. 1091–1098, 2001.

[2] M. F. Fast, S. Nill, J. L. Bedford, and U. Oelfke, "Dynamic tumor tracking using the Elekta Agility MLC," *Med. Phys.*, vol. 41, no. 11, 2014, Art. no. 111719.

[3] T. Depuydt et al., "Geometric accuracy of a novel gimbals based radiation therapy tumor tracking system," *Radiotherapy Oncol.*, vol. 98, no. 3, pp. 365–372, 2011.

[4] L. V. Romaguera et al., "Prediction of in-plane organ deformation during free-breathing radiotherapy via discriminative spatial transformer networks," *Med. image Anal.*, vol. 64, 2020, Art. no. 101754.

[5] D. Ruan, J. A. Fessler, and J. Balter, "Real-time prediction of respiratory motion based on local regression methods," *Phys. Med. Biol.*, vol. 52, no. 23, 2007, Art. no. 7137.

[6] A. Sawant, S. Dieterich, M. Svatos, and P. Keall, "Failure mode and effect analysis-based quality assurance for dynamic MLC tracking systems," *Med. Phys.*, vol. 37, no. 12, pp. 6466–6479, 2010.

[7] M. Hoogeman, J.-B. Prévost, J. Nuyttens, J. Pöll, P. Levendag, and B. Heijmen, "Clinical accuracy of the respiratory tumor tracking system of the cyberknife: Assessment by analysis of log files," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 74, no. 1, pp. 297–303, 2009.

[8] E. W. Pepin, H. Wu, Y. Zhang, and B. Lord, "Correlation and prediction uncertainties in the cyberknife synchrony respiratory tracking system," *Med. Phys.*, vol. 38, no. 7, pp. 4036–4044, 2011.

[9] A. Krauss, S. Nill, M. Tacke, and U. Oelfke, "Electromagnetic real-time tumor position monitoring and dynamic multileaf collimator tracking using a siemens 160 MLC: Geometric and dosimetric accuracy of an integrated system," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 79, no. 2, pp. 579–587, 2011.

[10] J. L. Bedford et al., "Effect of mlc tracking latency on conformal volumetric modulated ARC therapy (VMAT) plans in 4D stereotactic lung treatment," *Radiotherapy Oncol.*, vol. 117, no. 3, pp. 491–495, 2015.

[11] W. Kilby, J. Dooley, G. Kuduvalli, S. Sayeh, and C. Maurer Jr, "The cyberknife robotic radiosurgery system in 2010," *Technol. Cancer Res. Treat.*, vol. 9, no. 5, pp. 433–452, 2010.

[12] H. Shirato et al., "Feasibility of insertion/implantation of 2.0-mm-diameter gold internal fiducial markers for precise setup and real-time tumor tracking in radiotherapy," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 56, no. 1, pp. 240–247, 2003.

[13] R. Hirai, Y. Sakata, A. Tanizawa, and S. Mori, "Real-time linear fiducial marker tracking in respiratory-gated radiotherapy with a deep neural network," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 108, no. 3, pp. e259–e260, 2020.

[14] H. Lin, W. Zou, T. Li, S. J. Feigenberg, B.-K. K. Teo, and L. Dong, "A super-learner model for tumor motion prediction and management in radiation therapy: Development and feasibility evaluation," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, 2019.

[15] A. E. Bourque, J.-F. Carrier, É. Filion, and S. Bedwani, "A particle filter motion prediction algorithm based on an autoregressive model for real-time MRI-guided radiotherapy of lung cancer," *Biomed. Phys. Eng. Exp.*, vol. 3, no. 3, 2017, Art. no. 035001.

[16] D. Putra, O. Haas, J. Mills, and K. Bumham, "Prediction of tumour motion using interacting multiple model filter," in *Proc. IET 3rd Int. Conf. On Adv. Med. Signal Inf. Process.*, 2006, pp. 1–4.

[17] S. Hong and W. Bukhari, "Real-time prediction of respiratory motion using a cascade structure of an extended Kalman filter and support vector regression," *Phys. Med. Biol.*, vol. 59, no. 13, pp. 3555–3573, 2014.

[18] H. D. Hesar and M. Mohebbi, "An adaptive Kalman filter bank for ECG denoising," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 1, pp. 13–21, Jan. 2021.

[19] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[20] R. Wang, X. Liang, X. Zhu, and Y. Xie, "A feasibility of respiration prediction based on deep bi-LSTM for real-time tumor tracking," *IEEE Access*, vol. 6, pp. 51262–51268, 2018.

[21] H. Lin et al., "Towards real-time respiratory motion prediction based on long short-term memory neural networks," *Phys. Med. Biol.*, vol. 64, no. 8, 2019, Art. no. 0 85010.

[22] S. Yu, J. Wang, J. Liu, R. Sun, S. Kuang, and L. Sun, "Rapid prediction of respiratory motion based on bidirectional gated recurrent unit network," *IEEE Access*, vol. 8, pp. 49424–49435, 2020.

[23] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3633–3642.

[24] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[25] S. Li et al., "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 5243–5253, 2019.

[26] P. Jafari et al., "In-vivo lung biomechanical modeling for effective tumor motion tracking in external beam radiation therapy," *Comput. Biol. Med.*, vol. 130, 2021, Art. no. 104231.

[27] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.

[28] K. McCall and R. Jeraj, "Dual-component model of respiratory motion based on the periodic autoregressive moving average (periodic ARMA) method," *Phys. Med. Biol.*, vol. 52, no. 12, 2007, Art. no. 3455.

[29] S. Vedam et al., "Predicting respiratory motion for four-dimensional radiotherapy," *Med. Phys.*, vol. 31, no. 8, pp. 2274–2283, 2004.

[30] H. Wu et al., "A finite state model for respiratory motion analysis in image guided radiation therapy," *Phys. Med. Biol.*, vol. 49, no. 23, 2004, Art. no. 5357.

[31] M. Çetinkaya, S. A. Akgün, A. M. Erkmen, and İ. Erkmen, "Exact Kalman filtering of respiratory motion," in *Proc. IEEE 6th Int. Conf. Control Eng. Inf. Technol.,* 2018, pp. 1–6.

[32] F. Ernst, R. Dürichen, A. Schlaefer, and A. Schweikard, "Evaluating and comparing algorithms for respiratory motion prediction," *Phys. Med. Biol.*, vol. 58, no. 11, 2013, Art. no. 3911.

[33] X. Bao, W. Gao, D. Xiao, J. Wang, and F. Jia, "Bayesian model-based liver respiration motion prediction and evaluation using single-cycle and double-cycle 4D CT images," in *Proc. IEEE Int. Conf. Med. Imag. Phys. Eng.*, 2019, pp. 1–6.

[34] A. Jöhl et al., "Performance comparison of prediction filters for respiratory motion tracking in radiotherapy," *Med. Phys.*, vol. 47, no. 2, pp. 643–650, 2020.

[35] M. Seregni et al., "Real-time tumor tracking with an artificial neural networks-based method: A feasibility study," *Physica Medica*, vol. 29, no. 1, pp. 48–59, 2013.

[36] W. Sun et al., "Respiratory signal prediction based on adaptive boosting and multi-layer perceptron neural network," *Phys. Med. Biol.*, vol. 62, no. 17, 2017, Art. no. 6822.

[37] T. P. Teo et al., "Feasibility of predicting tumor motion using online data acquired during treatment and a generalized neural network optimized with offline patient tumor trajectories," *Med. Phys.*, vol. 45, no. 2, pp. 830–845, 2018.

[38] W. Sun, M. Jiang, G. Chen, and T. You, "A comparison of adaptive boosting algorithms for the respiratory signal prediction," in *Proc. 12th Int. Congr. Image Signal Process. BioMed. Eng. Informat.*, 2019, pp. 1–5.

[39] P. Chang et al., "Real-time respiratory tumor motion prediction based on a temporal convolutional neural network: Prediction model development study," *J. Med. Internet Res.*, vol. 23, no. 8, 2021, Art. no. e27235.

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

[41] G. Wang et al., "Real-time liver tracking algorithm based on LSTM and SVR networks for use in surface-guided radiation therapy," *Radiat. Oncol.*, vol. 16, no. 1, pp. 1–12, 2021.

[42] Z. Shen et al., "Su-f-j-141: Real-time position management (RPM) system as a valuable tool to predict tumor position deviation in sbrt lung and liver patients with breath hold using active breathing coordinator (ABC)," *Med. Phys.*, vol. 43, no. 6Part11, pp. 3440–3440, 2016.

[43] S. Yu et al., "Correlated skin surface and tumor motion modeling for treatment planning in robotic radiosurgery," *Front. Neurorobot.*, vol. 14, 2020, Art. no. 582385.

[44] F. Ernst, *Compensating for Quasi-Periodic Motion in Robotic Radiosurgery*. Berlin, Germany: Springer, 2011.

[45] R. Dürichen, T. Wissel, F. Ernst, and A. Schweikard, "Respiratory motion compensation with relevance vector machines," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*. Springer, 2013, pp. 108–115.

[46] H. Peng, L. Deng, Z. Xia, Y. Xie, and J. Xiong, "Unmarked external breathing motion tracking based on b-spline elastic registration," in *Proc. Int. Conf. Intell. Robot. Appl.*, 2021, pp. 71–81.

[47] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *Int. Stat. Review/Revue Inte. de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[48] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybernet.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[49] J. W. Tukey et al., *Exploratory Data Analysis*, vol. 2. Reading, MA, USA: Sage, 1977.

[50] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15. [Online]. Available: https://arxiv.org/abs/1412.6980

[52] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[53] S. Sayeh, J. Wang, W. T. Main, W. Kilby, and C. R. Maurer, *Respiratory Motion Tracking for Robotic Radiosurgery*. Berlin, Heidelberg, Germany: Springer, 2007.

[54] Y. Tsunashima et al., "Correlation between the respiratory waveform measured using a respiratory sensor and 3D tumor motion in gated radiotherapy," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 60, no. 3, pp. 951–958, 2004.

[55] S. A. Oh, J. W. Yea, S. K. Kim, and J. W. Park, "Optimal gating window for respiratory-gated radiotherapy with real-time position management and respiration guiding system for liver cancer treatment," *Sci. Rep.*, vol. 9, no. 1, pp. 1–6, 2019.

[56] A. Balasubramanian, R. Shamsuddin, B. Prabhakaran, and A. Sawant, "Predictive modeling of respiratory tumor motion for real-time prediction of baseline shifts," *Phys. Med. Biol.*, vol. 62, no. 5, 2017, Art. no. 1791.

[57] I. Y. Ha, M. Wilms, H. Handels, and M. P. Heinrich, "Model-based sparse-to-dense image registration for realtime respiratory motion estimation in image-guided interventions," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 2, pp. 302–310, Feb. 2019.