applied sciences

MDPI

Article

LSTM-Based Transformer for Transfer Passenger Flow Forecasting between Transportation Integrated Hubs in Urban Agglomeration

Min Yue [1] and Shuhong Ma [1,2,*]

[1] Department of Transportation Engineering, College of Transportation Engineering, Chang'an University, Xi'an 710064, China
[2] Key Laboratory of Transport Industry of Management, Control and Cycle Repair Technology for Traffic Network Facilities in Ecological Security Barrier Area, Chang'an University, Xi'an 710064, China
* Correspondence: msh@chd.edu.cn

Abstract: A crucial component of multimodal transportation networks and long-distance travel chains is the forecasting of transfer passenger flow between integrated hubs in urban agglomerations, particularly during periods of high passenger flow or unusual weather. Deep learning is better suited to managing massive amounts of traffic data and predicting extended time series. In order to solve the problem of gradient explosion or gradient disappearance that recurrent neural networks are prone to when dealing with long time sequences, this study used a transformer prediction model to estimate short-term transfer passenger flow between two integrated hubs in an urban agglomeration and a long short-term memory network to incorporate previous historical data. The experimental analysis uses two sets of transfer passenger data from the Beijing-Tianjin-Hebei urban agglomeration, collected every 30 min in May 2021 on the transfer corridors between an airport and a high-speed railway station. The findings demonstrate the high adaptability and good performance of the suggested model in passenger flow forecasting. The suggested model and forecasting outcomes assist management in making capacity adjustments in time to correspond with changes, enhance the effectiveness of multimodal transportation systems in urban agglomerations and significantly enhance the service of long-distance multimodal passenger travel.

Keywords: transfer passenger flow forecasting; transformer; deep learning; long short-term memory; multimodal transportation system

Citation: Yue, M.; Ma, S. LSTM-Based Transformer for Transfer Passenger Flow Forecasting between Transportation Integrated Hubs in Urban Agglomeration. *Appl. Sci.* **2023**, *13*, 637. https://doi.org/10.3390/app13010637

Academic Editors: Xue Yang and Luliang Tang

Received: 27 November 2022
Revised: 28 December 2022
Accepted: 28 December 2022
Published: 3 January 2023

1. Introduction

In China, the rate of urbanization has increased, which has sped up the movement of people, information and resources across metropolises. Long-distance multimodal travel inside urban agglomerations or between them has increased significantly; this has increased requirements for multimodal transportation systems' operating effectiveness. The study of multimodal transport modes [1] and the sharing rate of each mode in long-distance travel [2] related to transportation networks is a priority, and the study of long-distance travel mode choice, transfer choices and travel costs related to individual choices is another [3–5].A major source of concern for policymakers is the forecast of passenger flow between significant cities [6] as well as the demand for travel by air [7] and high-speed railway [8] in megacities.

Numerous techniques based on conventional statistics (such as ARIMA [9], gray prediction models [10], etc.), primarily developed from neural networks (such as ANN [11], CNN [12], and DNN [13]), have been extensively used in past studies to solve prediction issues in the field of traffic. The ability of prediction algorithms to handle vast amounts of data is becoming more and more important as data collecting capabilities advance. More forms of data, such as cab trajectory data, cell phone signaling data, travel application

data, etc., can be employed for traffic forecast analysis. Researchers have attempted to use deep learning techniques to promote the application of neural network-type models in massive amounts of traffic data. These techniques include recurrent neural networks [14], long short-term memory networks [15] and gated recurrent units [16] and various hybrid models [17–19]. These researchers have achieved rich results, including long, medium and short term prediction of traffic station flow [20], transportation mode flow [18,21] and traffic networks [19] based on massive amounts of traffic data, based on the widely used LSTM, GRU and other algorithms. Lately, the transformer algorithm has been better used in traffic timeseries prediction [22]; it can train the model by removing the spatiotemporal characteristics of traffic data [23], and also helps with the dependence issue within long series data processing [24].

Transfer between different modes of transportation is one of the critical links in the long-distance multimodal travel chain of urban agglomeration, which is frequently impacted by exceptional events, such as severe weather, holiday festival gatherings and unexpected line interruptions. Interchange difficulties may result in travelers missing the next leg of their trip or trip suspension for multimodal travelers. Urban agglomerations' multimodal transportation management and operation organizations must accurately estimate and track real-time transfer passenger volumes inside key transfer routes. The accompanying huge passenger flow response plan and the transfer passenger forecast results used together prior to the occurrence of extreme weather or mega-events can improve the operational effectiveness of the urban cluster multimodal transportation system. For the purpose of developing short-term projections of transfer passengers between integrated hubs within urban agglomerations, the Transformer's classical encoder–decoder framework is taken into consideration in this study. Additionally, LSTM is employed to pre-process passenger flow statistics, addressing the issues with long time series data dependency and extracting passenger flow features for simple encoder recognition. The decoder maps the features to the prediction sequence using the attention method to obtain the future value. The analyzed model's overall structure is shown in Figure 1, which calls for historical data for practical training purposes. It forecasts future traffic passenger tendencies by entering real-time data on transfer passengers between particular hubs. The prediction findings can be utilized as the foundation for evaluating trends in transfer passenger aggregation and implementing the monitoring feature of cooperative multimodal transportation system operation in the urban agglomeration. The model suggested in this study can accept input series in parallel and without the idea of the time step, in contrast to classic forecasting models that typically pass input time series data one after another.

By creating a deep learning-based transformer architecture, this study assesses the suitability of an LSTM-based transformer for short-time transfer passenger flow prediction. The Beijing-Tianjin-Hebei urban agglomeration's air-rail transfer corridor between Beijing Capital Airport and Beijing South Railway Station was chosen as the study's research object. It used a deep learning model transformer to perform a short-term prediction of the transfer passenger flow using analysis of data from cell phone signals. In the Beijing-Tianjin-Hebei urban agglomeration, the forecasting results are used as a reference base for monitoring and managing multimodal traffic operations by placing the historical data. The following are the research contributions made by this work.

(1) For short-term prediction of inter-hub transfer passengers in the urban agglomeration, the transformer framework based on attention mechanisms is applied. A more valuable and adaptable prediction method is created for large-scale time series prediction of interchange passengers with notable temporal characteristics.
(2) To resolve the long-time dependence in the time series prediction process and to turn the historical traffic passenger sequences into vectors that the transformer can recognize, the LSTM is utilized to pre-process the historical transfer passenger data.
(3) The passenger transfer between two significant passenger hubs in the Beijing-Tianjin-Hebei urban agglomeration is forecasted using the LSTM-based transformer based on

10–23 May 2021, and the forecast results serve as a reference point for increasing the operational effectiveness of the multimodal transportation system.

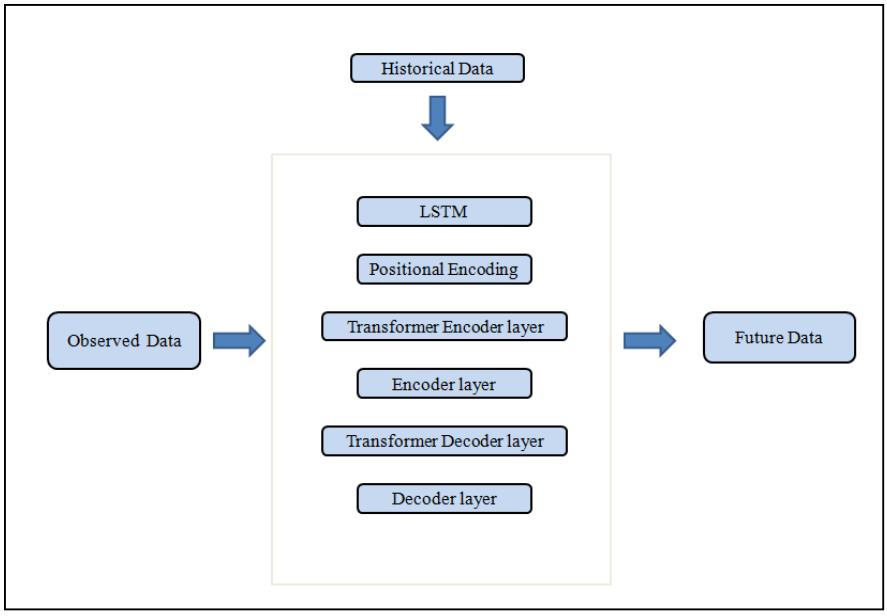

Figure 1. Illustration of the overall idea of the model.

The following structure describes how this essay is set up. Section 2 introduces the pertinent theories, models, research developments and applications of deep learning to temporal prediction and traffic modeling. The LSTM-based transformer model architecture and techniques, and the metrics that can be used to assess the model's ability to predict outcomes are introduced in Section 3. Section 4 predicts the transfer passengers' short time between Beijing Capital Airport and Beijing South Railway Station (railway station) in the urban agglomeration of Beijing-Tianjin-Hebei and analyzes the findings. The research discussion and conclusion are presented in Section 5.

2. Related Work

2.1. Recurrent Neural Networks, Long Short-Term Memory and Gated Recurrent Units

Recent years have seen a significant increase in the literature on deep learning algorithms used in traffic prediction research. These algorithms, which are based on the RNN model and its variant models LSTM and GRU, can effectively address the shortcomings of conventional methods, such as their low efficiency for processing large amounts of data. The advanced characteristics, constraints and prediction outcomes of RNN and LSTM algorithms in the traffic domain are first introduced in this section.

Recurrent neural networks, or RNNs, are frequently used to predict time series in a variety of disciplines, including economics, medicine, meteorology, engineering, etc. Recurrent neural networks outperform conventional time series prediction models in many time series studies [25–27]. When dealing with scenarios with variable-length temporal inputs or outputs and context dependencies, variational models of neural networks such as CNNs and DNNs typically perform less well. The fundamental principle of recurrent neural networks is to extend the data along the time axis such that each moment of data corresponds to a single neural unit, allowing the outcome of one moment to be carried over to the next, which is more efficient for solving problems with lengthy input–output and context-dependent relationships. It performs better when predicting continuous data with long input–output edges and context dependencies than other neural networks. RNNs, which differ from feedforward neural networks in that they use association rules to link historical processing data with current data, have attained widespread use and are well liked in time series research. The application of recurrent neural networks for

the prediction of longer data sequences frequently suffers from optimization challenges and long training time problems that occur in general deep networks and are prone to gradient disappearance or gradient explosion problems. This is because RNNs are unable to learn linked information when the gap between the relevant information and the needed information becomes very large.

The gradient disappearance or gradient explosion problem can be solved using GRU or LSTM as a typical method. Sepp Hochreiter and JiirgenSchrnidhuber's proposal for the LSTM [28], a particular type of recurrent neural network in deep learning, has been studied for 24 years. Temporal data can be processed using LSTM, which is widely utilized in intelligent applications, including automated speech recognition and natural language processing [29,30]. Compared with the original recurrent neural network, LSTM can handle extended sequence data, making it the most common RNN variation at the moment. LSTM is frequently employed in forecasting research projects in fields such as traffic flow and speed [15,31]. The benefit of the LSTM is that it uses its distinct forgetting and memory mechanisms to selectively access and store useful information. The forgetting gate, input gate and output gate are three distinct "gate" structures that govern which new information is added to the control memory state, which old information is discarded and which new state is produced. Although RNNs and its variations can address the issue of traditional statistical methods' and neural network models' inefficiency in managing vast amounts of data when predicting, they still need to be improved when handling complex time series or time series of different durations. The encoder–decoder framework based on attention mechanism, which draws on the expertise of applying deep learning in natural language processing, may be able to address this issue.

2.2. Attention Mechanism and Transformer

Even though LSTM can resolve the long sequence dependence issue, the model must receive each input data set separately, which is not practicable for complex time series computing. The latest attention mechanism significantly improves the above issue. The attention mechanism gained notoriety in 2014 when Google Mind published "Recurrent Models of Visual Attention"; this work used the RNN model and incorporated the attention mechanism for picture categorization [32]. The Seq2Seq&attention methodology for machine translation was first applied to the NLP area in 2015 when Bahdanau published the article "Neural Machine Translation by Jointly Learning to Align and Translate" [33]. The 2017 publication "Attention is All You Need" by the Google machine translation team ultimately rejected network architectures such as RNN and CNN. Machine translation jobs only used the attention mechanism [34]. Better outcomes were obtained, and the attention mechanism immediately gained the interest of many academics. The reduced performance of long sentence translation in natural language was the primary problem that the attention mechanism was intended to address. The attention mechanism has undergone ongoing improvement and has been applied extensively for temporal prediction [34–36]. The attention process illustrates two key points. The weights and weighted averages of changeable context vectors are calculated using the softmax function. The attention mechanism has been widely employed in many models, including LSTM, Seq2Seq, the encoder–decoder framework, etc. because it is an effective method of obtaining information that enables the encoder to actively seek relevant information at each step and temporarily ignore irrelevant information. Zhu developed a multichannel LSTM neural network that uses an attention layer to connect model outputs to input sequences in order to further increase prediction accuracy [37]. Kondo also used a sequence-to-sequence ("Seq2Seq") model to predict influenza epidemics; the findings revealed that their method performed better than ARIMA and LSTM-based models [38].

The Seq2Seq model's basic premise is to read an input sentence using a recurrent neural network, compress the information of the entire sentence into a fixed-dimensional encoder and then read this encoder using a second recurrent neural network to "de-compress" it into a sentence in the target language. The recurrent neural network that reads the input

sentence and decompresses the sentence in the target language is referred to as the encoder and decoder, respectively, in Cho's 2014 proposal of the encoder-decoder architecture, based on Seq2Seq [39]. The widely used encoder–decoder structure allows for selecting various recurrent neural networks per various requirements and usage scenarios. Transformer's foundation is a framework for encoders and decoders. Transformer, a multi-head self-attention mechanism developed by Google and based on the encoder–decoder, was widely adopted in natural language processing after it was submitted in 2017. The position of the input language is encoded using the encoder then converted into a vector by the decoder, utilizing the attention method. Transformer has recently started to be used for time series forecasting. It largely relies on its multi-head attention mechanism and layer stacking to capture dynamic hierarchical patterns in time series data, and to predict lengthy time series of challenging situations. Similarly to natural language processing, traffic timing issues research must concentrate on networks' temporal and spatial aspects. As a result, the transformer model has some promise for use in network and traffic flow prediction.

Transform models may be used to address long-term correlation and multidimensional dynamic dependence of temporal traffic data, since the data in the traffic domain has greater dimensional features than natural language data. The transformer was used to estimate the future by using hierarchical data to learn the dynamic and spatiotemporal properties of traffic [40]. To address the shortcomings of the existing traffic prediction research's algorithm for capturing spatial data, another study developed the temporal spatial transformer framework [23]. To determine traffic network's geographical and temporal properties, Xu uses the transformer model, which serves as the foundation for forecasting [22].

3. Methods

3.1. Methodological Architecture

In this section, the transformer architecture—using LSTM as the temporal embedding module—will be developed for transfer passenger forecasting of transfer corridors connecting integrated transportation hubs in urban agglomerations. The process of estimating the future values of a series from a time series of historical values is known as time series forecasting. To precisely predict time series using deep learning approaches, a set of historical values are first used to train a model by extracting temporal features, and then the model is used to predict future values of the time series. The historical time series is retrieved as input using an encoder for feature extraction, and turned into a vector coupled to a decoder through a self-attentive mechanism when the Transformer model is employed for time series prediction. In order to forecast future values in an autoregressive fashion, the decoder is used to locate the vector and concentrate on the most valuable portion of the previous data. After deciding on the LSTM model to extract temporal features from historical input data, we built this study's architecture in accordance with the original transformer architecture, as shown in Figure 2.

The suggested model's components based on the architecture will be discussed in turn, and numerous measurements of prediction accuracy are offered.

3.2. Temporal Embedding with LSTM

When used in the NPL domain, the transformer model produces good results, and its potential application to time series prediction is still being investigated. Time series forecasting requires data embedding before the encoder of the initial transformer model can identify the historical data directly. Data embedding turns the historical data into vectors. This study opts to extract the temporal features from the historical data through the LSTM for better time series modeling work, since the temporal features of the traffic passenger flow data are easily overlooked by conventional conversion of historical information. Because LSTM is frequently used in time series prediction models and can ignore insignificant features in historical time series while keeping critical information, it was chosen for this project.

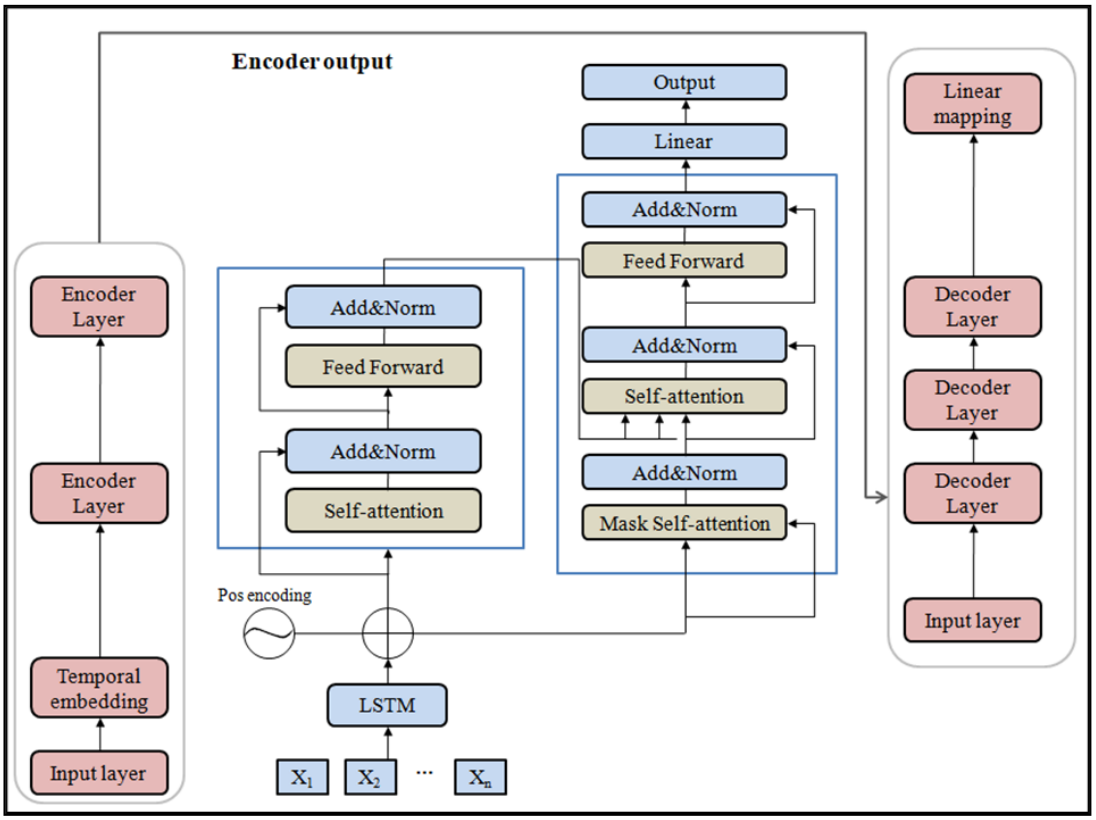

Figure 2. LSTM-based transformer model architecture.

The forgetting gate, memory gate and output gate are the three components of the LSTM, which comprises three pieces in total. The formula expression and the possible functions are as follows:

Step1: Select the data you want to erase using the forgetting process. The values on each dimension are translated using the sigmoid function in this phase by reading the last state output h_{t-1} and the current input x_t. The information on the dimension represented by the values close to 1 will be maintained, while the values around 0 will be forgotten.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f \quad (1)$$

Step2: The input mechanism chooses which portion of the information can be input after the recurrent neural network discards the data that does not need to be recalled. This information is needed to supplement the important data. In this stage, a single new candidate value that will be added to the state sequence is created using the tanh layer.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i \quad (2)$$

$$\widetilde{C}_t = \tan h(W_c[h_{t-1}, x_t] + b_C \quad (3)$$

Step3: The output for the current instant through the output gate must be made after the calculation to acquire the new state. The output mechanism chooses what value should be output and then presents a filtered version, after adding the new candidate values produced by the memory mechanism to the additional state information.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \quad (4)$$

$$y_t = \sigma(W_y[h_{t-1}, x_t] + b_y \quad (5)$$

$$h_t = y_t * \tan h(C_t) \quad (6)$$

3.3. Encoder

An input layer, a position encoding layer and an encoder layer make up the stack that is the encoder. As seen in Figure 2, the historical time series is first embedded by an LSTM to create the input layer, after which the location is encoded using the sin-cos rule. Location and relative position are equally as important to traffic passenger prediction as natural language understanding. The historical time series are disrupted when the time series are changed, which also has an impact on the model's training effect. One of the basic techniques employed by Transformer is the sine and cosine function method of position encoding. The vector formed after position encoding is input to the two encoder layers. To encode the sequential information in the time series, the components of the input vector and the position-encoded vector are combined. The self-attention layer and a fully linked feedforward sublayer, both followed by a normalization layer, are the same components of both encoders.

By linearly transforming the encoding vectors into three matrices, Q, K, and V, which stand for *Query*, *Key*, and *Value*, the self-attentive layer aids in the capture of the characteristics. In essence, the attention mechanism employs Q and K to compute "attention weights" before utilizing those weights to weight the sum of V.

$$\text{Attention}(\text{Q},\ \text{K},\ \text{V}) = \text{softmax}\left(\frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}_\text{k}}}\right)\text{V} \tag{7}$$

The self-attention mechanism is used to derive the attention matrix which depicts the probability distribution of the attention weights. Then, V is weighted using the attention matrix and normalized using softmax. The completely linked feedforward sublayer can help to mine the hidden characteristics between distinct nodes, and can further enhance the model at its site.

3.4. Decoder

The decoder architecture used in this work is based on the transformer classical model and consists of an input layer, three identical decoder levels, and an output layer. Each decoder layer comprises a feedforward neural network, masked multi-head attention and multi-head attention. Residual links are created between each sublayer, and layer normalization is then applied to stabilize the gradients and aid in model training. To ensure that the prediction of time series data points relies only on prior data points and ignores non-local features for improved extraction of local features, masked multi-head attention is positioned between the input of the decoder and the desired output. The input layer begins with the final bit of encoder input data, converts decoder input to a vector of d_{model} dimensions and then converts the last decoder output to target time series.

3.5. Evaluation

According to Equation (8), the assessment statistic for this study is the mean absolute percentage error (MAPE).

$$\text{MAPE} = \frac{100\%}{\text{n}}\sum_{\text{i}=1}^{\text{n}}\left|\frac{\hat{\text{y}}_\text{i} - \text{y}_\text{i}}{\text{y}_\text{i}}\right| \tag{8}$$

Predicted value: $\hat{\text{y}} = \{\hat{\text{y}}_1, \hat{\text{y}}_2, \ldots\ \ldots, \hat{\text{y}}_\text{n}\}$

True value: $\text{y} = \{\text{y}_1, \text{y}_2, \ldots\ \ldots, \text{y}_\text{n}\}$

The metric calculation yields results in the range of (0, +), with a MAPE of 0% denoting a flawless model and a MAPE of more than 100% denoting a subpar model.

4. Experiments and Results

4.1. Experiments Environment and Data Description

The training method is facilitated by an RTX3080 station running CUDA 10.1. The computer's i9-10900x station CPU has tensor core GPUs covering a 128 GB GPU memory.

Python coding was performed on the open-source PyTorch machine learning toolkit. For the experimental study, the passenger transfer data for two weeks (passengers every 30 min, 672 data items (48 times/day × 14 days) in one direction for transfer corridors) between Capital Airport and Beijing South Railway Station inside the Beijing-Tianjin-Hebei urban agglomeration are chosen.

The core megacity, Beijing, is part of the Beijing-Tianjin-Hebei urban agglomeration, which also consists of two sizable cities (Shijiazhuang and Tianjin) and 110 million permanent residents. It's one of China's most significant urban agglomerations. It is crucial in reducing Beijing's non-capital functions, developing the ties between Beijing and Tianjin and improving the province of Hebei's all-around capacity. Within the Beijing-Tianjin-Hebei urban agglomeration, the transfer passenger flow between Beijing Capital International Airport and Beijing South Railway Station is chosen for analysis in the study. The Beijing Capital International Airport's flight passenger throughput exceeded 100 million, placing it at the top of the country's list in 2018 and 2019.The Beijing South Railway Station is a significant integrated transportation hub that connects high-speed railroads, regular railroads, urban railroads, urban public transportation and other modes of transportation. The locations of Beijing Capital Airport and Beijing South Railway Station on the urban agglomeration are shown in detail in Figure 3, with the former being in the northeast of Beijing and the latter being on the southern side. A transfer corridor between air-rail intermodal hubs is represented in Figure 3 by a blue line with double arrows on it.

The original balance of intra-city transportation has been upset by the expansion of air-rail intermodal passenger traffic, which has led to a new need for transit between numerous hubs within the urban agglomeration. In addition to the composite travel chain of various transportation mode combinations that are available for users to pass between hubs, the air-rail intermodal passengers between the Capital Airport and Beijing South Railway Station displaya variety of complex travel behaviors, using the bus, subway and taxis. According to Figure 4, which shows the various access options available to cross-hub passengers in urban multimodal transportation networks, one significant factor is the growth in the number of passengers between hubs and the heterogeneity of individual passenger characteristics and travel attributes.

As a result, it is challenging to determine the transfer passenger flow data between the two hubs using operational data from the airport or high-speed railway station. Instead, we may estimate inter-hub transfer passenger traffic utilizing data from cell phone signaling by employing new data sources. Cell phone users can be identified as inter-hub transfer passengers by users who pass through both Beijing Capital Airport and Beijing South Railway Station at the chosen time, because they passively connect to the base station network at least once every 30 min without actively using the network. The Unicom cell phone signaling data from the National Key Research and Development Program 2018YFB1601300 were chosen for this study, and following data cleaning and sample enlargement, we obtained the data of the interchange passenger flow between Beijing Capital Airport and Beijing South Railway Station from 10–23 May 2021. Every 30 minwas selected as a statistical period to count the number of inter-hub transfer passengers from 10–23 May 2021, taking into account the base station's capacity to capture the users' cell phone signaling data. The trend is depicted in Figure 5a, where the horizontal coordinates represent the statistical dates and the vertical coordinates represent the number of transfer passengers (in persons) in each statistical period, with 48 statistics placed between each pair of adjacent dates. For the selected 14 days and 676 sets (48 sets of data per day) of one-way passenger flow statistics for inter-hub transfer corridors are available. As an illustration, Figure 5b depicts the transfer passengers counted every 30 min on one of the days (10 May).

In Figure 5a, line 1, shown as a yellow line, depicts the transfer passenger flow between Beijing South Railway Station (BSRS) and Capital Airport (BCA) every 30 min between May 10 and May 23, 2021, line 2 (the blue line) depicts the 30 min transfer between Beijing South Railway Station (BSRS) and Beijing Capital Airport (BCA) From 10–23 May 2021.

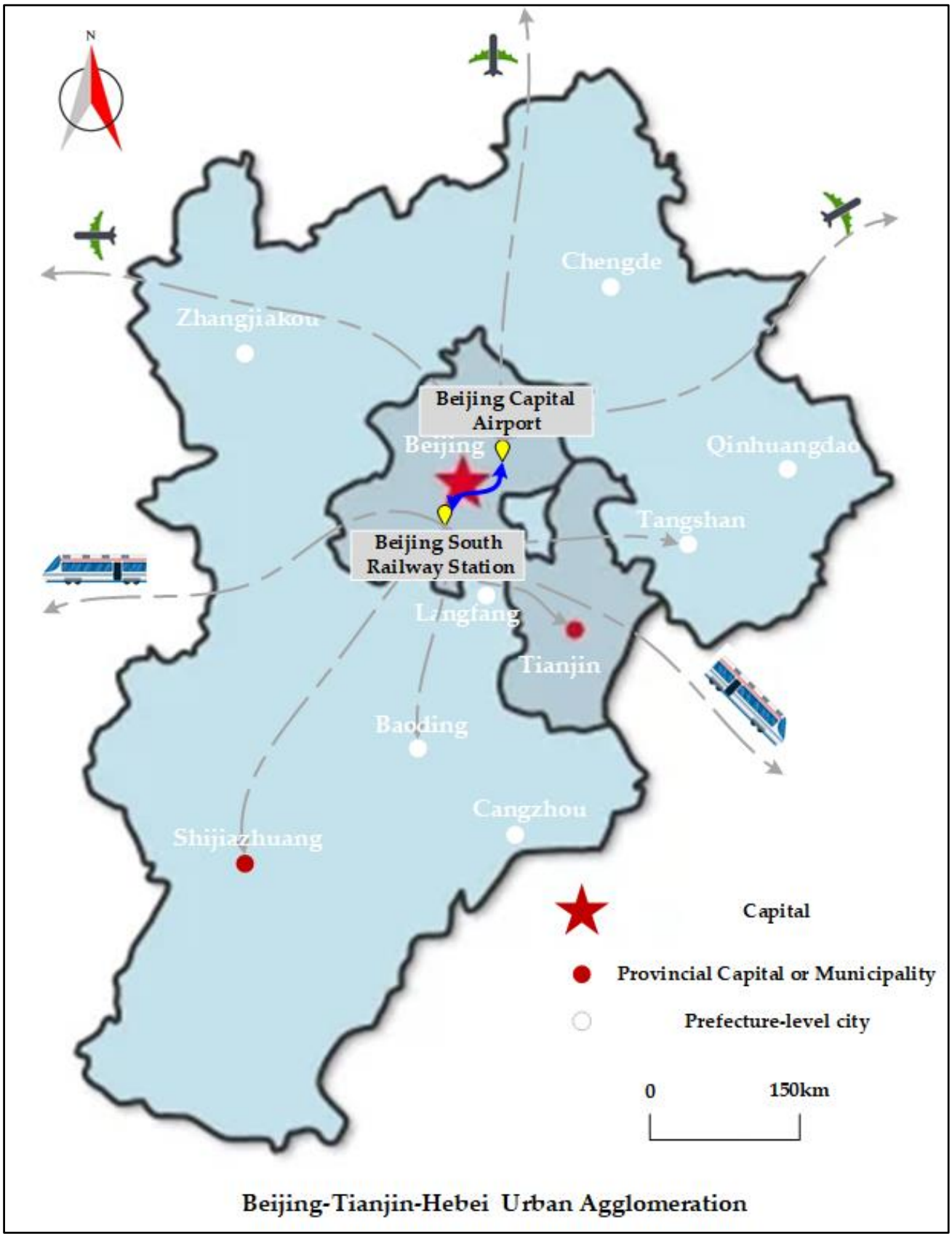

Figure 3. The location of Beijing Capital Airport and Beijing South Railway Station in the Beijing-Tianjin-Hebei urban agglomeration.

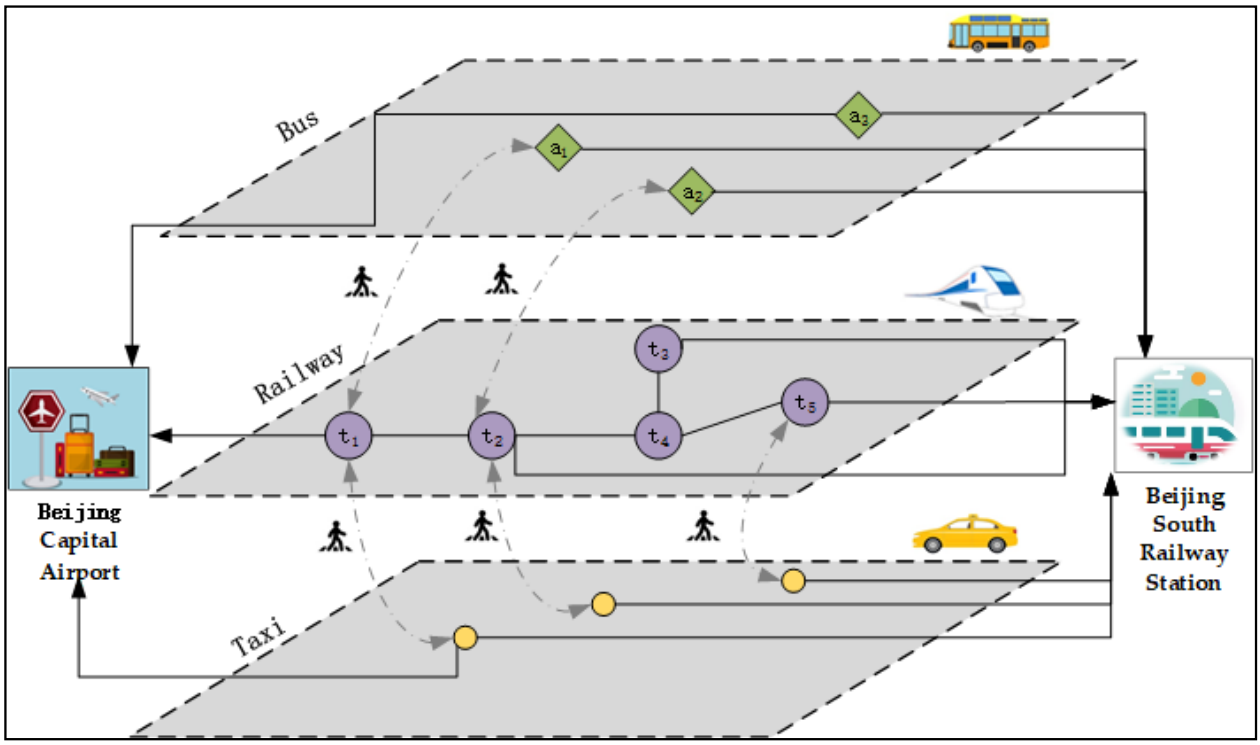

Figure 4. Air-rail transfer line between the Beijing Capital Airport and Beijing Nan railway station.

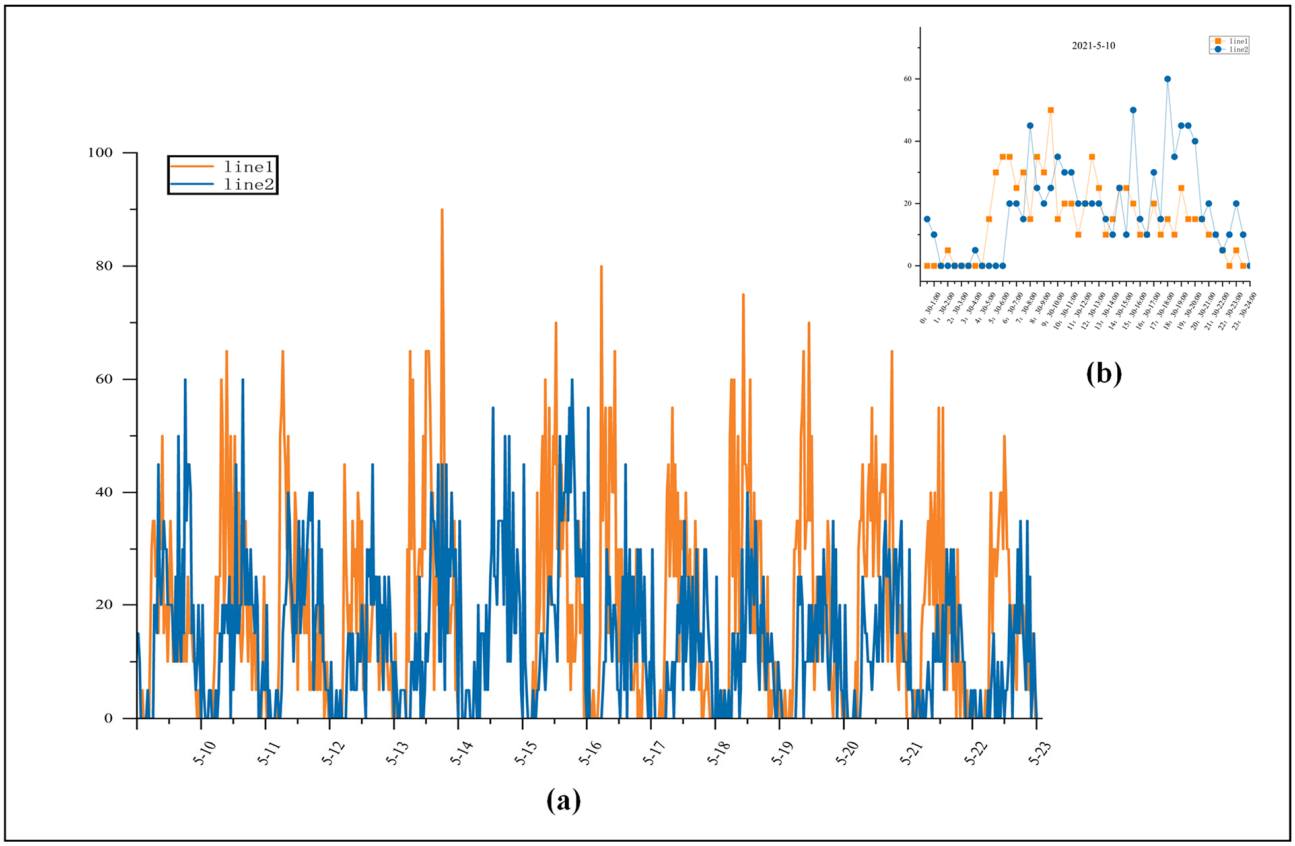

Figure 5. Transfer passenger flow between BCA and BSRS from 10 May 2021 to 23 May 2021 (**a**) Transfer passenger flow trend from 10 May 2021 to 23 May 2021. (**b**) The transfer passengers counted every 30 min on 10 May 2021.

4.2. Data Pro-Processing and Training

LSTM pre-processes the historical time series to mask the uninteresting features, as mentioned in Section 3.1.The relevant embedding vectors are given the position information by the sin-cos function, which selects the output vector sequence. The following parameters were first chosen for the transformer model's encoder construction in order to guarantee the model's top performance in light of various examples of temporal prediction using the transformer model:

$$feather\_size = 8,$$

$$num\_layers = 2,$$

$$dropout = 0.5;$$

The following parameters are chosen to aid in the smooth operation of the model for the decoder component of the transformer:

$$feather\_size = 8,$$

$$num\_layers = 3,$$

$$dropout = 0.5;$$

During the training process, the learning rate is set to 0.0001, and the computational loss and optimization effect are observed with the random descent of the MSEloss loss function. In the transformer model training process, the choice of batch_size determines the number of times that the trained samples are into the computational method. The decline and oscillation magnitude of MSEloss are observed when trying batch_size of 4,6,8,16, respectively (Figure 6). It is finally confirmed that the whole dataset is convergence when choosing batch_size = 16.

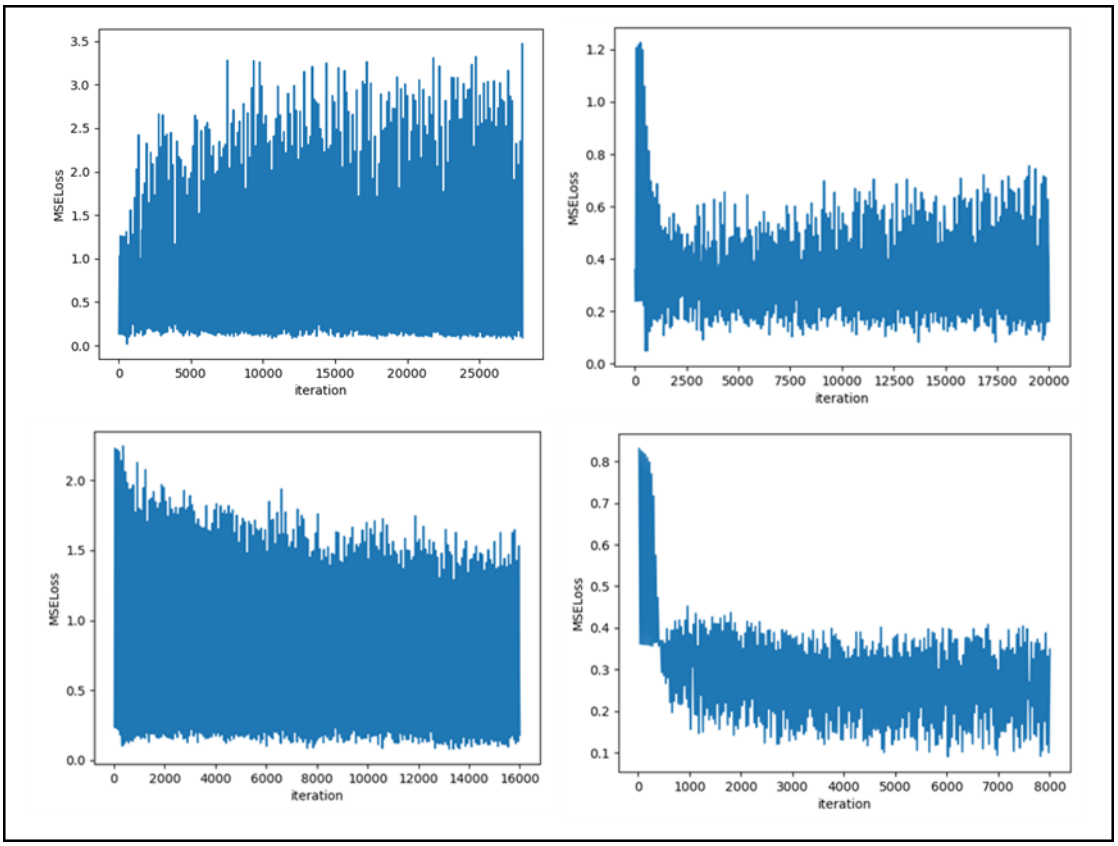

Figure 6. Training loss of different batch_size (4,6,8,16).

One epoch signifies the conclusion of the forwarding and backward training of the entire training set once during the training of a neural network. The complete training set must typically be run through the same neural network multiple times because one epoch is typically insufficient to process the entire training set in a neural network. Overfitting results from using too many epochs, whereas using too few epochs results in suboptimal training parameters. In this study, the batch size was set to 16, and we tested the Transformer training at epochs of 50, 100, 200 and 300 to obtain the MSEloss function decrease graph shown in Figure 7. It is clear that the convergence effect is satisfactory at epochs of 200 or 300.

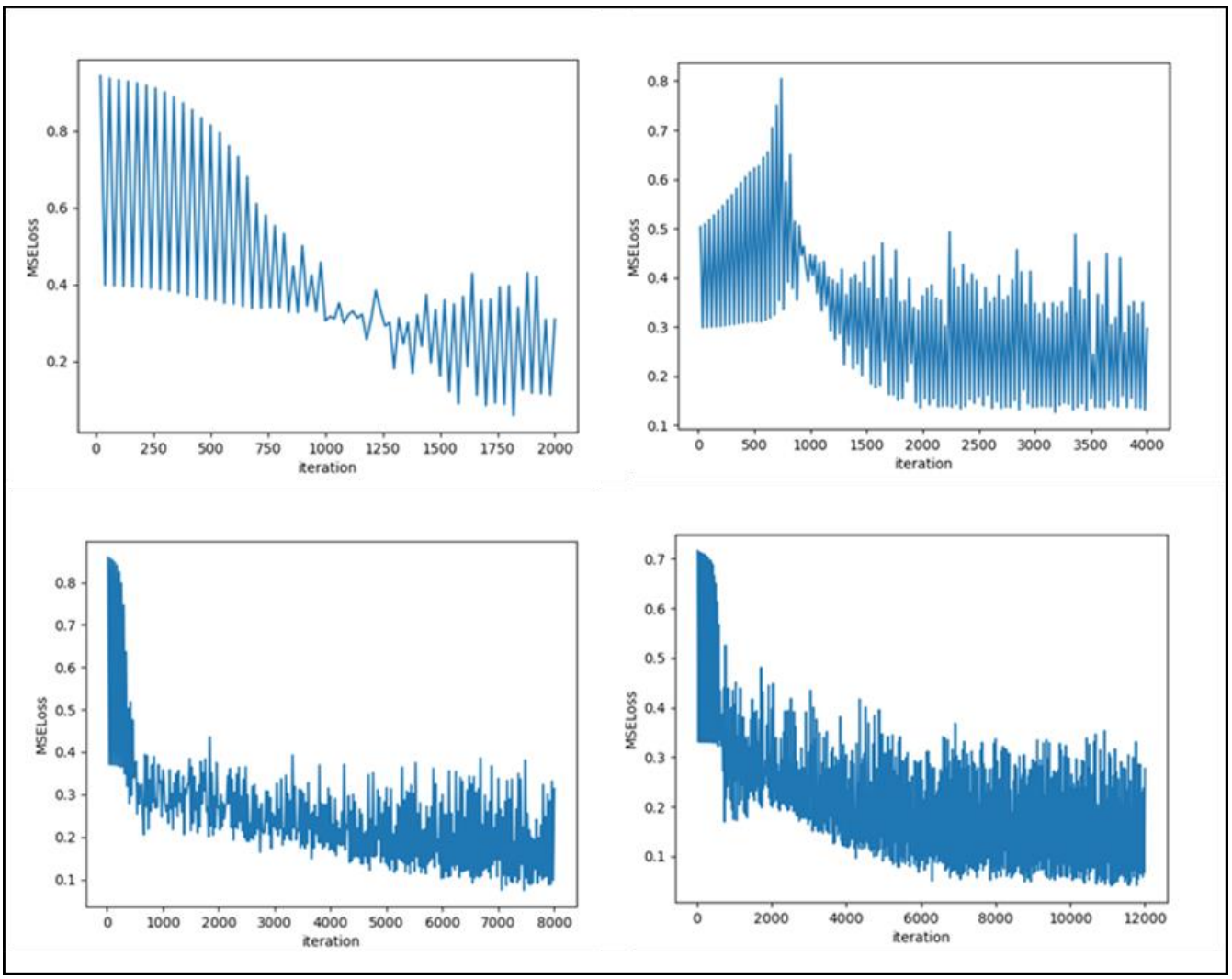

Figure 7. Training loss of different epochs (50,100,200,300).

4.3. Results

After multiple tries, the best results are produced by a decoder with three stacked layers and an encoder with two stacked layers. The following parameters are in accordance with the description information in the previous section:

$$Learning\ rate = 0.0001$$

$$Batch\_size = 16$$

$$Epoch = 200$$

By training the LSTM-based transformer model with the historical data from the previous 11 days, we conducted trials to see if it could predict the passenger transfer for the following three days. The results of the training and testing of the interchange corridors' transfer passenger flow from BCA to BSRS and from BSRS to BCA are displayed in Figures 8 and 9, respectively. The MSEloss functions that were used in the training process are shown in Figures 8a and 9a in this case, while Figures 8b and 9b depict the prediction of transfer passengers, with real values corresponding to 144 statistical periods (every 30 min) for the three days from 21 May 2021 to 23 May 2021. "0" designates the time period from 0:00 to 0:30 on 21 May 2021.

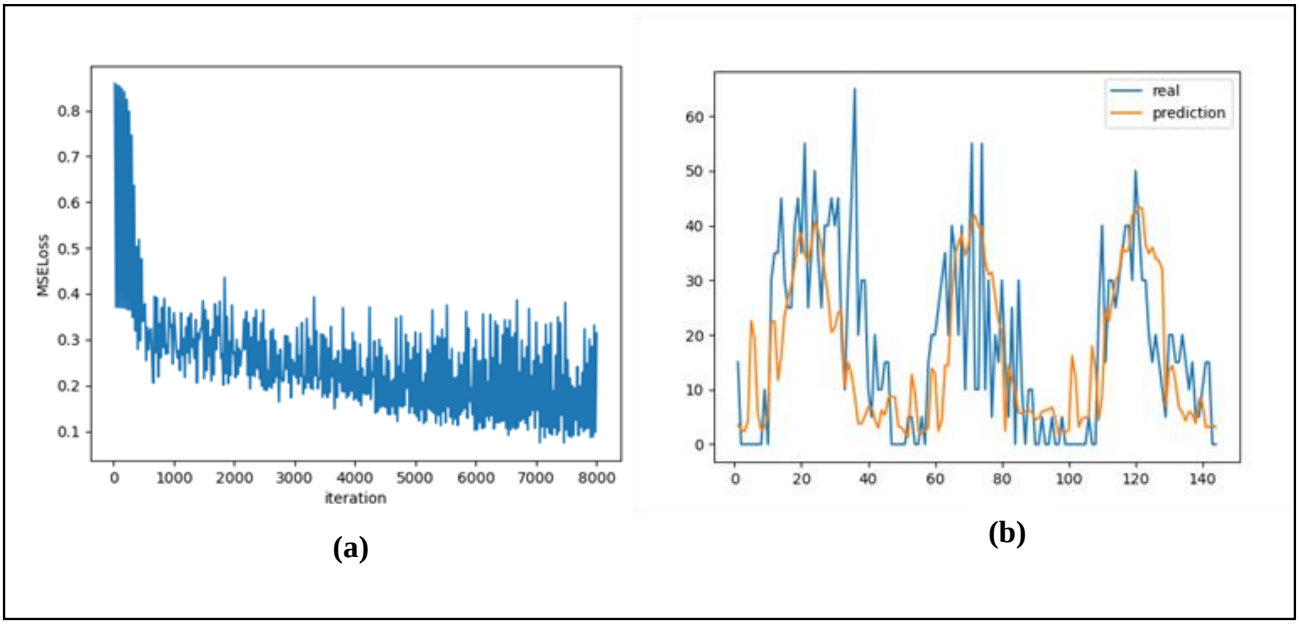

Figure 8. The transfer passenger flow forecasting results of transfer corridors from BCA to BSRS. (**a**) Training MSEloss function. (**b**) The forecasting results.

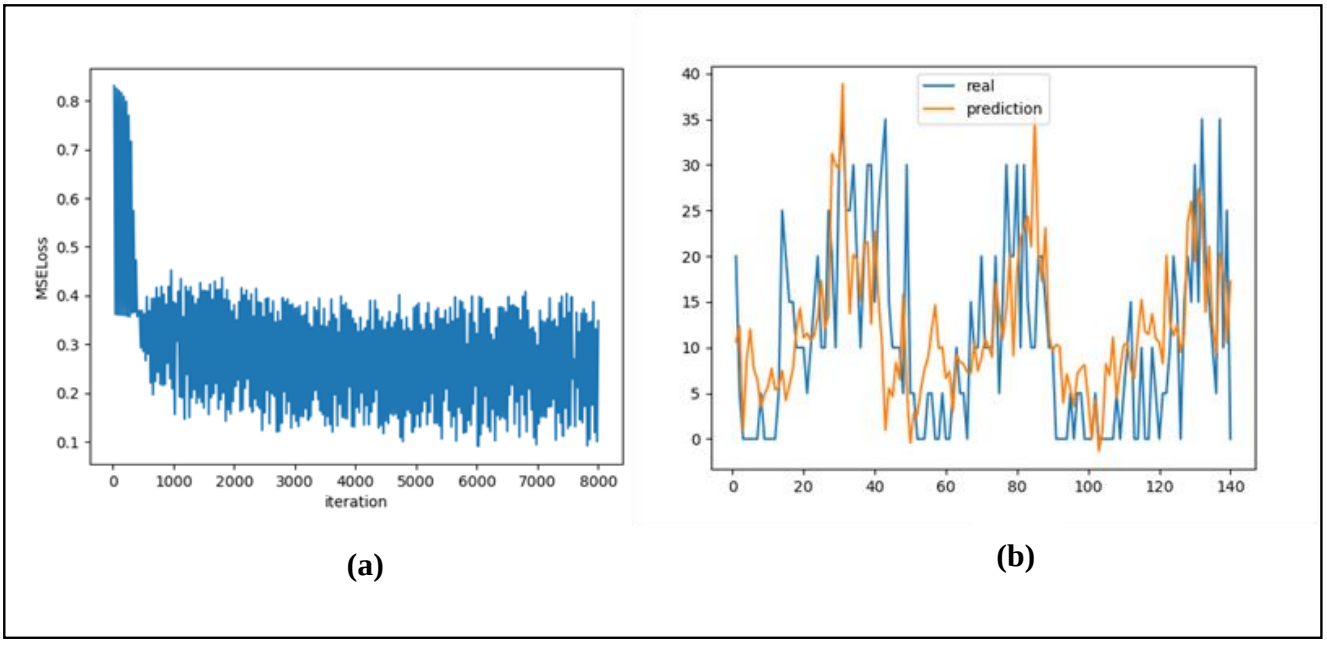

Figure 9. The transfer passenger flow forecasting results of transfer corridors from BSRS to BCA. (**a**) Training MSEloss function. (**b**) The forecasting results.

After 200 training epochs, using the historical data of the interchange passenger flow of the BAC-BSRS interchange from 10–20 May 2021, as the training set, it can be seen in Figure 8 that the MSEloss function (Figure 8a) is decelerating noticeably. However, the oscillation is still within the acceptable range. The test set's prediction vs. the actual values are displayed in Figure 8b. The MAPE between the anticipated and real values when combined with the outcomes of the code runs is 21.33%, demonstrating the model's effectiveness.

The drop in the MSEloss function (Figure 9a) is not as good as the training outcome of the prior interchange channel, as can be seen in Figure 8. Using the historical data of the interchange passenger flow of the BSRS-BAC transfer corridor from 10–20 May 2021, as the training set, and after 200 training epochs, it still exhibits a declining trend, and the oscillation is within the acceptable range. Figure 9b displays the test set's anticipated values in comparison with their actual values. The MAPE between the predicted and real values is 27.12% when combined with the code run results, showing that the model is effective.

5. Conclusions

Making forecasts of transfer passenger flow between significant passenger transportation hubs in urban agglomerations is essential to better serve urban agglomerations' integrated multimodal transportation operation. The data obtained are crucial for ensuring an efficient multimodal transportation system. Therefore, based on the conceptual framework and model structure depicted in Figures 1 and 2, this study suggests a transformer model based on LSTM. The suggested approach uses a self-attention mechanism to model the time series data and focuses more on capturing temporal aspects in traffic flow. By contrasting it with other time series forecasting techniques, the capacity of this deep learning system to learn complicated dependencies of different lengths from traffic time series data is increased (such as statistics and traditional neural networks). The consistent length of the input sequence is no longer required, making it easier to extend the application of the prediction method. To test the model performance, we chose the passenger transfer between two significant integrated passenger hubs (the Beijing Capital Airport and Beijing South Railway Station) in the Beijing-Tianjin-Hebei urban agglomeration as the experimental object, and we used cell phone signaling from 10–23 May 2021, as the base data. The time series forecasting analysis uses the extracted and extended data: data from 10–20 May 2021, model training and transfer flow projection from 21–23 May 2021, and actual date comparison. A powerful method for forecasting traffic flow data with temporal features is demonstrated by the LSTM-based transformer model, which obtains advanced results on the actual transfer flow dataset in both directions.

In addition to the short-term prediction of inter-hub transfer passenger flow, the method proposed in this study can also be used to predict more multimodal traffic network data in urban agglomerations, such as the prediction of transport volume in important passenger hubs and the prediction of different modes of passenger flow between significant cities. It can also be used to predict passenger flow in inter-hub transfer corridors in different states, especially when there are significant travel periods, such as peak travel seasons, significant holidays or when significant cultural or sporting events are happening in a location. Both multimodal transportation managers in urban agglomerations and users of multimodal transportation networks might benefit from the findings of projecting passenger flows along transfer corridors at particular times. The forecasted outcomes of inter-hub transfer passenger flow, from the manager's perspective, can assist management in planning for high passenger flows, considering expanding the capacity of specific transit modes based on the forecast data or adjusting the frequency of inter-hub rail operations to ensure that the hub can easily handle the influx of passengers during special occasions. Once the management combines the forecast results to optimize the capacity adjustment of the transfer corridors, it will be very beneficial for traveling passengers to ensure that long-distance multimodal travel chain users are in the process of travel mode switching smoothly and quickly to avoid travel disruption, cancellation or long delays caused by extraordinary events.

Last but not least, the simplified format required for the input sequence and the applicability of the approach to the prediction of nodes or channels within the traffic network demonstrate the progressive nature of the work of the transformer model proposed in the study. However, the outcomes of the transformer run in the experimental instance are not indisputably better than those of the LSTM and GRU models. In Selim's research [24] about the traffic flow forecasting method, the MAPE of the LSTM model achieved 12.37%, the MAPE of the GRU model reached 12.66% and the best forecast outcome's MAPE in this study was 21.33%. One of the study's research weaknesses is the dearth of historical passenger flow data, which contributed to the fact that the effect of MAPE in the tests was not very noteworthy. Due to confidentiality laws governing the data used in this study, only two weeks' worth of traffic passenger flow data along interchange corridors could be collected. When 30 min is chosen as the statistical period, the number of statistical periods included in the given period is only 676 sets (one-way), which is less for validating deep learning algorithms. Future studies could produce better outcomes, since larger sample sizes will make data easier to access during training and testing.

Additionally, due to the impact of COVID-19, the multimodal traffic flow for the Beijing-Tianjin-Hebei urban agglomeration during 2019–2021 has decreased compared with the pre–2019 period. The results obtained by selecting May 2021, a period less affected by the pandemic, are more in line with the standard operating conditions. In future research, we will continue to gather passenger flow data for inter-hub transfer corridors in urban agglomerations as COVID-19's effects progressively fade; we will train and test the model with more statistical data and continuously improve the model to obtain better results.

Author Contributions: Conceptualization, M.Y. and S.M.; methodology, M.Y.; formal analysis, M.Y.; investigation, M.Y.; resources, S.M.; data curation, M.Y.; writing—original draft preparation, M.Y.; writing—review and editing, M.Y; funding acquisition, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: The research is supported by the National Key Research and Development Program of China (2018YFB1601300).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Some or all data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request.

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

References

1. Gangyan, X.; Ruibing, Z.; Suxiu, X.; Xiaofei, K.; Xuan, Q. Personalized Multimodal Travel Service Design for sustainable intercity transport. *J. Clean. Prod.* **2021**, *308*, 127367.
2. Yan, H.; Huiming, Z. The intercity railway connections in China: A comparative analysis of high-speed train and conventional train services. *Transp. Policy* **2022**, *120*, 89–103.
3. Dorian, A.B. Individual, household, and urban form determinants of trip chaining of non-work travel in México City. *J. Trans. Geogr.* **2022**, *98*, 103227.
4. Min, Y.; Shu-hong, M.; Wei, Z.; Xi-fang, C. Estimation Markov Decision Process of Multimodal Trip Chain between Integrated Transportation Hubs in Urban Agglomeration Based on Generalized Cost. *J. Adv. Transp.* **2022**, *2022*, 5027133.
5. Wong, Y.Z.; Hensher, D.A.; Mulley, C. Mobility as a service (MaaS): Charting a future context. *Transp. Res. Part A Policy Pract.* **2019**, *131*, 5–19. [CrossRef]
6. Xiaowei, L.; Ruiyang, M.; Yanyong, G.; Wei, W.; Bin, Y.; Jun, C. Investigation of factors and their dynamic effects on intercity travel modes competition. *Travel Behav. Soc.* **2021**, *23*, 166–176.
7. Korkmaz, E.; Akgüngör, A.P. The forecasting of air transport passenger demands in Turkey by using novel meta-heuristic algorithms. *Concurr. Comp. Pract. Exp.* **2021**, *33*, e6263. [CrossRef]
8. Xie, M.Q.; Li, X.M.; Zhou, W.L.; Fu, Y.B. Forecasting the Short-Term Passenger Flow on High-Speed Railway with Neural Networks. *Comput. Intel. Neurosc.* **2014**, 375487. [CrossRef]
9. Wang, X.; Zhang, N.; Chen, Y.; Zhang, Y. Short-term forecasting of urban rail transit ridership based on ARIMA and wavelet decomposition. *Proc. AIP Conf.* **2018**, *1967*, 040025.

10. Li, J.W. Short-Time Passenger Volume Forecasting of Urban Rail Transit Based on Multiple Fusion. *Appl. Mech. Mater.* **2014**, *641*, 773–776. [CrossRef]
11. Alekseev, K.P.G.; Seixas, J.M. Forecasting the Air Transport Demand for Passengers with Neural Modelling. In Proceedings of the Brazilian Symposium on Neural Networks, Pernambuco, Brazil, 11–14 November 2002; pp. 86–91.
12. Ma, X.; Dai, Z.; He, Z.; Ma, J.; Wang, Y.; Wang, Y. Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction. *Sensors* **2017**, *17*, 818. [CrossRef] [PubMed]
13. Yarin, G.; Zoubin, G. A theoretically grounded application of dropout in recurrent neural networks. *NIPS* **2016**, *29*, 1019–1027.
14. Huang, W.; Song, G.; Hong, H.; Xie, K. Deep Architecture for Traffic Flow Prediction: Deep Belief Networks With Multitask Learning. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 2191–2201. [CrossRef]
15. Ma, X.; Tao, Z.; Wang, Y.; Yu, H.; Wang, Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transp. Res. Part C Emerg. Technol.* **2015**, *54*, 187–197. [CrossRef]
16. Rui, F.; Zuo, Z.; Li, L. Using LSTM and GRU neural network methods for traffic flow prediction. In Proceedings of the 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, China, 11–13 November 2016; pp. 324–328.
17. Xiao, Y.; Liu, J.J.; Hu, Y.; Wang, Y.; Lai, K.K.; Wang, S. A neuro-fuzzy combination model based on singular spectrum analysis for air transport demand forecasting. *J. Air Transp. Manag.* **2014**, *39*, 1–11. [CrossRef]
18. Jinlei, Z.; Feng, C.; Guo, Y. Multi-graph convolutional network for short-term passenger flow forecasting in urban rail transit. *IET Intel. Transp. Syst.* **2020**, *14*, 1210–1217.
19. Cui, Z.; Henrickson, K.; Ke, R.; Wang, Y. Traffic Graph Convolutional Recurrent Neural Network: A Deep Learning Framework for Network-Scale Traffic Learning and Forecasting. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 4883–4894. [CrossRef]
20. Li, L.; Wang, Y.; Zhong, G.; Zhang, J.; Ran, B. Short-to-medium Term Passenger Flow Forecasting for Metro Stations using a Hybrid Model. *KSCE J. Civ. Eng.* **2017**, *22*, 1937–1945. [CrossRef]
21. Zhizhen, L.; Hong, C. Short-Term Online Taxi-Hailing Demand Prediction Based on the Multimode Traffic Data in Metro Station Areas. *J. Transp. Eng. Part A Syst.* **2022**, *148*, 05022003.
22. Xu, M.; Dai, W.; Liu, C.; Gao, X.; Lin, W.; Qi, G.; Xiong, H. Spatial-Temporal Transformer Networks for Traffic Flow Forecasting. *arXiv* **2020**, arXiv:2001.02908.
23. Huaxiu, Y.; Xianfeng, T.; Hua, W.; Guanjie, Z.; Zhenhui, L. Revisiting Spatial-Temporal Similarity A Deep Learning Framework for Traffic Prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5668–5675.
24. Reza, S.; Ferreira, M.C.; Machado, J.J.M.; Tavares, J.M.R. A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural. *Expert Syst. Appl.* **2022**, *202*, 117275. [CrossRef]
25. Rangapuram, S.S.; Seeger, M.W.; Gasthaus, J.; Stella, L.; Wang, Y.; Januschowski, T. Deep state space models for time series forecasting. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), Montreal, QC, Canada, 3–8 December 2018; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 7796–7805.
26. Salinas, D.; Flunkert, V.; Gasthaus, T.J.; Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **2019**, *36*, 1181–1191. [CrossRef]
27. Wen, R.; Torkkola, K.; Narayanaswamy, B.; Madeka, D. A multi-horizon quantile recurrent forecaster. *arXiv* **2017**, arXiv:1711.11053.
28. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
29. Wen, T.H.; Gasic, M.; Mrksic, N.; Su, P.H.; Vandyke, D.; Young, S. Young, Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. *arXiv* **2015**, arXiv:1508.1745.
30. Graves, A.; Jaitly, N.; Mohamed, A.R. Hybrid speech recognition with deep bidirectional LSTM. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013; pp. 273–278.
31. Khan, Z.; Khan, S.M.; Dey, K.; Chowdhury, M. Development and Evaluation of Recurrent Neural Network-Based Models for Hourly Traffic Volume and Annual Average Daily Traffic Prediction. *Transp. Res. Rec. J. Transp. Res. Board* **2019**, *2673*, 489–503. [CrossRef]
32. Volodymyr, M.; Nicolas, H.; Alex, G. Recurrent Models of Visual Attention. *Adv. Neural Inf. Proces. Syst.* **2014**. [CrossRef]
33. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2015**, arXiv:1409.0473.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *NIPS* **2017**. [CrossRef]
35. Choi, E.; Bahadori, M.T.; Sun, J.; Kulas, J.; Schuetz, A.; Stewart, W. *Retain: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism*; NIPS: Barcelona, Spain, 2016; pp. 3504–3512.
36. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.X.; Yan, X. Enhancing the locality and breaking the memory bottleneck of Transformer on time series forecasting. *NeurIPS* **2019**. [CrossRef]
37. Zhu, X.; Fu, B.; Yang, Y.; Ma, Y.; Hao, J.; Chen, S.; Liu, S.; Li, T.; Liu, S.; Guo, W.; et al. Attention-based recurrent neural network for influenza epidemic prediction. *BMC Bioinform.* **2019**, *20*, 575. [CrossRef] [PubMed]
38. Kondo, K.; Ishikawa, A.; Kimura, M. Sequence to Sequence with Attention for Influenza Prevalence Prediction using Google Trends. In Proceedings of the 2019 3rd International Conference on Computational Biology and Bioinformatics, Nagoya, Japan, 17–19 October 2019. [CrossRef]

39. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734.
40. Haoyang, Y.; Xiaolei, M. Learning Dynamic and Hierarchical Traffic Spatiotemporal Features with Transformer. *IEEE Transact. Intell. Transp. Syst.* **2021**, *23*, 11.