Abstract—The short-term passenger flow prediction of the urban rail transit system is of great significance for traffic operation and management. The emerging deep learning-based models provide effective methods to improve prediction accuracy. However, most of the existing models mainly predict the passenger flow on general weekdays or weekends. There are only few studies focusing on predicting the passenger flow on holidays, which is a significantly challenging task for traffic management because of its suddenness and irregularity. To this end, we propose a deep learning-based model named Spatial-Temporal Attention Fusion Network (STAFN) comprising a novel Multi-Graph Attention Network (MGATN), a Conv-Attention Block, and Feature Fusion Block for short-term passenger flow prediction on holidays. The MGATN is applied to extract the complex spatial dependencies of passenger flow dynamically and the conv-attention block is applied to extract the temporal dependencies of passenger flow from global and local perspectives. Moreover, in addition to the historical passenger flow data, the social media data, which has been proven that they can effectively reflect the evolution trend of passenger flow under events, are also fused into the feature fusion block of STAFN. The STAFN is tested on two large-scale urban rail transit AFC datasets from China on the New Year’s holiday, and the prediction performance of the model are compared with that of several conventional prediction models. Results demonstrate its better robustness and advantages among benchmark methods, which can provide overwhelming support for practical applications of short-term passenger flow prediction on holidays.

Index Terms — Deep learning, Spatial-Temporal Attention Fusion Network, Social media data, Short-term holiday passenger flow prediction.

I. INTRODUCTION

The urban rail transit (URT) system has experienced rapid development in recent decades. As an important component in intelligent URT systems, short-term passenger flow prediction has been extensively studied nowadays. It is a critically significant task because leveraging the results of short-term passenger flow prediction, passengers can better schedule their travel plans and operators can take corresponding measures to provide high-level services. However, it is also a critically challenging problem especially on holidays because the passenger flow contain complicated spatiotemporal characteristics and they are easily affected under specific scenarios. For example, during holidays, the passenger flow often varies significantly and the regularities of passenger flow on holidays are extremely different from weekdays, as shown in Figure 1. Therefore, how to conduct an accurate prediction of passenger flow, especially on holidays remains a necessary and challenging task.

Figure 1 Inbound Passenger Flow around 2021 New Year’s Week

To solve this problem, some scholars have studied short-term passenger flow prediction on holidays. For instance, Chen and Liang (2015) proposed a machine learning-based method which hybridized the support vector regression model with an adaptive genetic algorithm to predict the holiday daily tourist flow. Additionally, considering that the support vector machine can deal with complex nonlinear characteristics, Liu and Yao (2017) proposed a modified support vector machine to predict passenger flow on holidays. Moreover, Xie and Sun (2020) analyzed the spatiotemporal characteristics of holiday passenger flow and then established a deep learning model based on the modified backpropagation neural network (BPNN) to predict the passenger flow on holidays. These predictive models fill the gap in holiday passenger flow prediction. However, there are also several shortcomings as follows. First, most of these models ignore the fact that there is a limited sample size of holiday passenger flow. A limited sample size of passenger flow data on holidays generally increases the

This work was supported by the National Natural Science Foundation of China (Nos. 72210269, 71621001, 71825004).

Shuxin Zhang, Jinlei Zhang, Lixing Yang, Jiateng Yin and Ziyou Gao are with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing 100044. (e-mail: 22110249@bitu.edu.cn; zhangjinlei@bitu.edu.cn; lxyang@bitu.edu.cn; jyin@bitu.edu.cn; zyga@bitu.edu.cn) (Corresponding author: Jinlei Zhang.)
difficulty of accurately predicting the short-term holiday passenger flow. Second, most models just utilize conventional data such as passenger flow data and it is not adequate for accurate holiday passenger flow prediction. Only relying on passenger flow data cannot fully capture the spatiotemporal characteristics of passenger flow on holidays. Social media messages under specific scenarios have been proven that can provide reliable contacts for predicting traffic flow (Roy and Hasan et al., 2021). Hence, the motivation of this paper is to study how to fuse the limited holiday passenger flow data with social media to accurately predict the holiday passenger flow.

In this paper, we propose a deep learning model based on the GNNs and attention mechanism, namely Spatial-Temporal Attention Fusion Network (STAFN), for accurate network-scale short-term passenger flow prediction on holidays in the URT system. In addition to the conventional passenger flow data on holidays, social media data (microblogs data volumes) are also fused into the STAFN to fully capture the impact of related social media to passenger flow on holidays. We compare the model with other eleven prediction models such as ConvLSTM, Graph WaveNet (GWN), Diffusion Convolutional Recurrent Neural Network (DCRNN), and Transformer. Experimental results on two real-world urban rail transit AFC datasets from China show the superiority and the great robustness of the STAFN model. Specifically, the main contributions of this paper are as follows:

1. The passenger flow data and social media data (microblogs data volumes) related to holidays are fused into our model to improve the accuracy of short-term holiday passenger flow prediction.
2. We propose a deep-learning model called STAFN, which consists of the MGATN, the conv-attention block, and the feature fusion block for short-term passenger flow prediction in URT systems on holidays. The MGATN is applied to dynamically capture the complex spatial dependencies of stations in the subway network. The conv-attention block is presented because of its enhanced ability to capture temporal dependencies from global and local perspectives. The feature fusion block is used to fuse the social media and passenger flow data for learning the impact of related social media on the holiday passenger flow.
3. We utilize the passenger flow data on holidays for two consecutive years to fully capture the holiday characteristics, which solves the problem of poor prediction performance caused by the limited sample size of holiday passenger flow.
4. The advantages of STAFN are introduced by two network-wide urban rail transit AFC datasets from China on the New Year’s holiday from 2019 to 2021. Results show its favorable prediction performance in passenger flow prediction on holidays and the considerable robustness in different scenarios.

The remainder of this article is organized as follows: Section II reviews the literature about passenger flow prediction. Section III provides the problem definition. Section IV describes the details of the proposed STAFN model. In Section V, we evaluate the prediction performance of our model based on two real-world datasets. Finally, we make a conclusion of our paper in Section VI.

II. LITERATURE REVIEW

In this section, the short-term passenger flow prediction models, holiday passenger flow prediction, and the application of social media data are summarized.

A. Short-term Passenger flow prediction

In recent years, lots of short-term passenger flow prediction methods have been put forward (Liu, 2017; Zhang and Chen, 2019; Ma, 2021). Generally, the prediction methods can be divided into two categories, one is the regression model based on mathematical statistics, and the other is the machine learning-based model.

The general mathematical statistic methods include the Autoregressive Integrated Average model (ARIMA) and grey prediction model (Liu and Yang, 2016; Li and Xiang, 2020), etc. Zeng and Xu (2008) proposed a hybrid model that combined ARIMA and multilayer artificial neural networks for short-term traffic flow prediction. Kumar and Vanajakshi (2015) proposed a prediction scheme based on the seasonal ARIMA model for short-term prediction of traffic flow using only limited input data, which could overcome the problem of data availability. Ni et al.(2017) integrated linear regression and seasonal ARIMA to predict passenger flow under event occurrences. Considering that the nature of passenger flow prediction is a time series problem, it is believed that the key is to capture the spatiotemporal characteristics of passenger flow series, which is the weakness of statistical regression models like ARIMA. On the contrary, the machine learning-based models can better capture spatiotemporal features and thus outperform these classic models.

Machine learning-based models such as support vector regression (SVR) and random forest learning, which can fully capture nonlinear features and temporal characteristics of passenger flow data, are increasingly applied in traffic prediction (Leshem, 2007; Hong and Dong, 2011). Castro-Neto and Jeong (2009) presented an online SVR model for the prediction of short-term traffic flow under both typical and atypical conditions. Hu and Yan (2016) proposed a hybrid prediction method based on particle swarm optimization and SVR for short-term traffic flow prediction. However, most of these models only consider temporal characteristics and might not fully consider spatial correlations in the model formulation. Moreover, these models are generally verified on a single URT station and they are inapplicable for all of the stations in the entire network (Zhang and Chen, 2020).
As a branch of machine learning, deep learning-based model, such as BPNN (Zheng and Lee, 2006), CNN (Zhang and He, 2007; Zhang and Yu, 2019), LSTM (Ma and Tao et al., 2015), GCN (Yu and Lee, 2020), have also received lots of attention from industrial and academic in recent years. For instance, Yang and Chen (2019) proposed an improved model enhancing long-term features based on LSTM to capture long temporal dependence for URT passenger flow prediction. Ren et al. (2019) utilized a residual network (ResNet) based on CNN to model the spatiotemporal dependency to improve the traffic flow prediction accuracy. Wu and Tan (2016) presented a hybrid deep learning-based architecture combining CNN and LSTM to predict short-term traffic flow. To capture topological information, Xu et al. (2019) presented a spatiotemporal multi-graph convolution network (ST-MGCN) to conduct traffic demand prediction. Zhao and Song (2020) proposed a temporal graph convolutional network (T-GCN) model combining the GCN and the gated recurrent unit (GRU) for traffic flow prediction. Zhang and Chen (2021) developed a novel OD flow forecasting method that considered the unique characteristics of the URT system, which mainly consisted of a channel-wise attention mechanism and split CNN.

These short-term passenger flow prediction models have favorable performance on weekdays or weekends. However, due to the significant irregularity and fluctuation of passenger flow on holidays, these models might not fully capture spatiotemporal characteristics and holiday features, resulting in poor prediction performance on holidays. Hence, how to improve the accuracy of passenger flow prediction on holidays remains to be explored.

B. Holiday passenger flow prediction

To predict the holiday passenger flow accurately, much researches have been carried out. For example, Zeng and Sheng (2019) proposed a learning framework based on weighted knowledge transfer for daily peak load prediction during holidays. Luo and Li (2019) utilized a hybrid prediction model combining discrete Fourier transformer (DFT) with support vector regression (SVR) to extract common trends in the traffic flow for accurate holiday traffic flow prediction. Zhou et al. (2020) analyzed the characteristics of passenger flow on holidays and constructed a predictive model based on the support vector machine to realize the accurate prediction. To find a more reliable model under various conditions such as holidays, Zhang and Yao (2021) presented a hybrid deep spatiotemporal model combining convolutional neural network (CNN), gated recurrent unit (GRU), and convolutional long short-term memory (ConvLSTM) models. Considering the linear and nonlinear time series, Wen and Zhao (2022) proposed a decomposition-based predictive method with transfer learning, which has been proven that it is beneficial to improve the accuracy of passenger flow prediction on holidays.

Most of the forementioned models focus on the regularity of passenger flow on holidays. However, the sample size of historical passenger flow on holidays is too limited to reflect the specific pattern of holiday passenger flow, resulting in poor prediction performance. Hence, combining with other data that correlated with passenger flow may be a feasible way to improve the accuracy of prediction.

C. Social media application

In recent years, with the rapid development of social media, an increasing number of researchers have attempted to integrate social media data into the traffic prediction field. Considering that social media (Chaniotakis and Antoniou, 2015) can reflect users’ intentions, it is seen as an effective data source to be applied in passenger flow prediction, especially regarding special events. For instance, Ni et al. (2014) proposed a short-term traffic flow prediction model, incorporated with tweet features to predict incoming traffic flow before sports game events. A few years later (2017), they found that there existed a moderate positive correlation between passenger flow and the rates of social media posts. And they presented a parametric and convex optimization-based approach to predict subway passenger flow under event occurrences. Essien et al. (2020) proposed a deep learning prediction model based on Bi-directional LSTM and stacked autoencoder (SAE) that combined information extracted from tweet messages with traffic and weather information to improve predictive accuracy. Roy and Hasan (2021) utilized traffic sensors and Twitter data to predict traffic demand during hurricane evacuation. Xue and Liu (2022) firstly studied the multivariate disturbance that affects passenger flow of a station under event occurrences, and then they presented a three-stage deep learning model to model the disturbance of inbound flow from nearby stations and social media post trends.

The above literature shows that there is a great potential to apply social media to explore the comprehensive features of passenger flow under event occurrences, thus improving prediction accuracy. However, there is little research exploring how to apply social media to holiday passenger flow prediction in URT. In this paper, we present an approach combining historical passenger flow data and microblog data to predict passenger flow on holidays in URT.

III. PROBLEM DEFINITION

In this section, the problem definition is formulated. We first define several key parameters and then put forward the learning problem of short-term URT passenger flow prediction on holidays.

The purpose of this study is to use historical AFC data to predict the passenger flow of the URT system at the next time interval. The passenger flow data extracted from the AFC data can be counted and integrated at different time intervals (e.g. 10 minutes, 30 minutes). The time interval used in this study is 10 minutes.

Definition 1 (passenger flow matrix): A subway AFC data record includes the following information: a passenger’s card number, the arrival time of a passenger, the arrival station of a
passenger, the departure time of a passenger, and the departure station of a passenger.

Given passengers’ arrival information at a time period $t$ on station $n$, let $p_n(t)$ be the observed volumes on station $n$ during the $t$th time interval. We define the passenger flow matrix as follows:

$$p^{t-1} = \begin{pmatrix} 
    p_1(t-1) & p_1(t-2) & \cdots & p_1(t-q) \\
    p_2(t-1) & p_2(t-2) & \cdots & p_2(t-q) \\
    \vdots & \vdots & \ddots & \vdots \\
    p_n(t-1) & p_n(t-2) & \cdots & p_n(t-q) 
\end{pmatrix} \quad (1)$$

where $p^{t-1} \in R^{n \times q}$ denotes the observed inflow of the entire URT network during $t-1$ to $t$, $n$ denotes the number of stations in the whole URT systems, and $q$ denotes the maximum time steps during $t-1$ to $t$. In this paper, we use historical 12-time steps (after comparing the prediction performance of the model at different time steps) to predict passenger flow $Y_t$ at the next time step $t$.

**Definition 2 (social media feature matrix):** To capture the potential trend of passenger flow on holidays, a social media feature matrix $S^{t-1}$ is defined as follows:

$$S^{t-1} = \begin{pmatrix} 
    s_1(t-1) & s_1(t-2) & \cdots & s_1(t-q) \\
    s_2(t-1) & s_2(t-2) & \cdots & s_2(t-q) \\
    \vdots & \vdots & \ddots & \vdots \\
    s_n(t-1) & s_n(t-2) & \cdots & s_n(t-q) 
\end{pmatrix} \quad (2)$$

where $S^{t-1} \in R^{n \times q}$ denotes the holiday-related microblog volumes during $t-1$ to $t$, the other parameters are the same as the passenger flow matrix.

**Definition 3 (multiple traffic graphs):** To fully express the multiple spatial relationships of each station in the whole subway network, the adjacent graph, the functional similarity graph, and the OD correlation graph are introduced as $G_a = (S, E_a, W_a)$, $G_c = (S, E_c, W_c)$, and $G_e = (S, E_e, W_e)$, respectively. $S = \{s_1, s_2, \ldots, s_n\}$ denotes the set of stations in the whole network. Note that the relationships between stations on different graphs are different, we define $E_a$, $E_c$, and $E_e$ are the edge sets of different graphs. For each graph, we assign a specific weight to each edge $W_k(k = a, c, e) \in R^{N \times N}$, and $W_k(i, j)$ denotes the weight coefficient between the station $i$ and station $j$.

1) Adjacent Graph: $G_a$ is directly built according to the physical topology between the stations of the whole URT network. The edge $E_a(i, j)$ between station $i$ and station $j$ is constructed, if these two stations are connected in the real world. Meanwhile, if there exists $E_a(i, j)$ between station $i$ and station $j$, $A(i, j) = 1$, or else $A(i, j) = 0$. The common definition of the adjacent-based weight between station $i$ and station $j$ is formulated as:

$$W_a(i, j) = \frac{A(i, j)}{\sum_{k=1}^{k \in A(i, k)}} \quad \text{for } i, j = 1, 2, \ldots, N \quad (3)$$

2) Functional Similarity Graph: In addition to the physical topology between stations, the functional similarity of the stations is also critical to model the spatial dependencies. Although some stations are far apart, they may be related to each other because of similar functions (such as commuter stations, and business stations). Thus, we take the functional similarity into account to further capture the spatial dependencies. Specifically, the similarity weight $\mathcal{C}(i, j)$ between station $i$ and station $j$ is computed with Pearson Correlation Coefficient:

$$S(i, j) = \frac{\overline{E}(p[i, p[j]) - \overline{E}(p[i]) \overline{E}(p[j])}{\sigma(p[i]) \sigma(p[j])} \quad (4)$$

where $p[i] \in R^{t \times q}$ denotes the passenger flow of station $i$ during the $t$th time interval, and $q$ denotes the time steps during the $t$th time interval. $E(\cdot)$ denotes the mathematical expectation, and $\sigma(\cdot)$ denotes the mathematical variance.

Given similarity weight matrix $S \in R^{n \times n}$, we select some stations with high similarity weight to construct edges $E_e$. We preset a similarity threshold $\overline{W}_e$ to determine the stations as follows:

$$S(i, j) = \begin{cases} 
    S(i, j), & \text{if } S(i, j) > \overline{W}_e \\
    0, & \text{otherwise} 
\end{cases} \quad (5)$$

Based on the corrected similarity weight matrix $\overline{C}$, we adopt row-normalization to this matrix for easily training. The normalized matrix is calculated as follows.

$$W_e(i, j) = \frac{\overline{S}(i, j)}{\sum_{i, j} \overline{S}(i, j)}, \text{for } i, j = 1, 2, \ldots, N \quad (6)$$

3) OD Correlation Graph: $G_c$ is constructed by the origin-destination distribution of passenger flow. We first introduce the OD distribution ratio between station $i$ and station $j$ as follows.

$$C(i, j) = \frac{D(i, j)}{\sum_{k=1}^{k \in D(i, k)}}, i, j = 1, \ldots, N \quad (7)$$

where $D(i, j)$ denotes the number of passenger flow from station $i$ to station $j$. Then we select some stations with high distribution ratio by comparing to a preset distribution ratio $\overline{W}_c$.

$$C(i, j) = \begin{cases} 
    C(i, j), & \text{if } C(i, j) > \overline{W}_c \\
    0, & \text{otherwise} 
\end{cases} \quad (8)$$

We also apply row-normalization to this matrix to build the OD correlation weight matrix as follows.

$$W_c(i, j) = \frac{\overline{C}(i, j)}{\sum_{k=1}^{k \in \overline{C}(i, k)}}, \text{for } i, j = 1, 2, \ldots, N \quad (9)$$

Thus, the passenger flow prediction in the whole subway network during holidays can be formulated as follows.

**Problem:** At time interval $t-1$, all AFC transaction records, the multiple graph $G_a$, $G_c$, and $G_e$, and the microblog data related to the specific holiday are available. Therefore, the historical passenger flow matrix $P^{t-1}$, the social media feature matrix $S^{t-1}$, and multiple spatial features can be extracted, and be used to predict the network passenger flow $Y_t$ at the next time step $t$. Thus, the problem can be defined as follows.

$$Y_t = f(P^{t-1}, S^{t-1}, G_a, G_c, G_e) \quad (10)$$

where $f$ indicates the model to be learned during the training process.
IV. METHODOLOGY

In this section, we specifically describe our proposed STAFN model. The framework of STAFN will be presented first. Then, the three components of STAFN named the Multi-Graph Attention Network, the Conv-Attention Layer, and the Feature Fusion Block will be introduced respectively.

A. Spatial-Temporal Attention Fusion Network

The main idea of our model is to directly model the complex spatial and temporal dependencies with the attention mechanism dynamically. The Spatial-Temporal Attention Fusion Network (STAFN) framework is shown in Figure 2, which mainly consists of three modules, named Multi-Graph Attention Network (MGATN), Conv-Attention Block, and Feature Fusion Block, respectively. Specifically, MGATN takes into consideration of multiple graphs to dynamically capture complex spatial dependencies according to the input. Conv-Attention Block models the temporal dependencies of passenger flow from global and local perspectives. To ensure effective training as the model goes deeper, the residual block is applied for the output of the conv-attention block. In order to learn the impact of holidays on the passenger flow, we develop a novel feature fusion block to fuse the related social media and holiday passenger flow, so that explore the temporal features from related social media to reflect the trend of holiday passenger flow over time. Finally, we project the feature fusion matrix into a fully connected layer to predict the passenger flow $Y_t$ in the next time step $t$.

![Figure 2. The overview of the STAFN](image)

B. Multi-Graph Attention network

With the powerful ability to capture spatial correlation and topological information of the graph, Graph Convolutional Network (GCN) (Kipf and Welling, 2016) has attracted more and more attention in the traffic prediction field recently. However, the conventional GCNs always focus on a predefined static graph based on physical topologies, which leads to several limitations in learning spatial dependencies. On the one hand, the traffic network is not static, it will change over time, which means the spatial dependencies between nodes change dynamically. On the other hand, these models always only learn a single spatial feature and cannot model the multiple spatial dependencies. Therefore, in addition to the physical topologies, we also take into consideration of the functional similarity and the OD correlation to develop a novel Multi-Graph Attention Network (MGATN) for multiple dynamic spatial dependencies learning. Figure 3 shows the architecture of MGATN.

The significance of the attention mechanism has been studied extensively in previous research. It can dynamically calculate the attention weight of inter-nodes to represent their relationships. Hence, in MGATN, we first leverage the idea of attention mechanism (Vaswani and Shazeer et. al., 2017) to compute the dynamic attention weight matrix of inter-stations as follows.

$$W_d = \text{Softmax}(\frac{p^t (p^t)^T}{\sqrt{d_p}}) \in R^{N \times N}$$

(11)

where $p^t \in R^{N \times TS}$ denotes the passenger flow during the $t$-th time interval and $d_p$ denotes the dimension of $P$. The calculation process can dynamically adjust the attention weight of inter-stations according to the input.

Assume the dynamic attention weight matrix $W_d$ is available, we need to incorporate prior knowledge into MGATN to ensure the stability and interpretability of MGATN. In this paper, we introduce three graphs, named Adjacent Graph $G_a$, Functional Similarity Graph $G_s$, and OD Correlation Graph $G_c$, respectively to express multiple spatial attributes of inter-stations. Owing to the great performance of the GCN proposed by Kipf et al. (2016) , we use the GCN with the first approximation of Chebyhev spectral filter in our model. Let $\tilde{G} \in R^{N \times N}$ denotes specific normalized graph with self-loop, $P \in R^{N \times TS \times d_{in}}$ denotes the input features, $Z \in R^{N \times TS \times d_{out}}$ denotes the output features, and $W \in R^{d_{in} \times d_{out}}$ denotes the model parameter matrix. Our graph attention network can be defined as follows.

$$Z = (\tilde{G} + W_d)PW$$

(12)

Given the multiple output feature matrices $Z_a, Z_s, \text{ and } Z_c$, we fuse these matrices by concatenation operation and finally use linear transformation to generate the passenger feature matrix with multiple spatial dependencies $Z_{all}$.

$$Z_{all} = \text{Linear(concat}(Z_a, Z_s, Z_c))$$

(13)

Note that the inflow data on holidays are significantly different from that on normal weekdays, and there is no obvious correlation between the holiday passenger flow and the weekday passenger flow. Thus, it is inappropriate to use the common three patterns: the real-time pattern, the daily pattern, and the weekly pattern to predict the passenger flow on holidays. In this paper, we only utilize the passenger flow with the real-time pattern. Suppose the time step is $ts$, and the passenger flow of the current time $t$ is to be predicted. The real-time pattern can be described as $P_{\text{real}} = (P_{t-\text{ts}}, P_{t-\text{ts} + 1}, \ldots, P_{t-1})$, a segment of historical time series adjacent to the predicting period. The passenger flow in the neighboring time period will directly influence the passenger flow in the next time period. The output passenger flow feature matrix of the real-time pattern $Z_{all}$ are integrated and then input into the modified Transformer layer.
C. Conv-Attention layer

Transformer architecture (Vaswani and Shazeer et al., 2017) was originally proposed for Natural Language Processing (NLP) and has achieved great success. Its significant module attention mechanism can effectively learn the attention weight of the sequence data to highlight the influence of important features on the output. Considering that traffic flow data is a typical time series data, it is of great significance to capture the temporal characteristics for traffic flow prediction. Therefore, in this paper, we apply the attention mechanism to capture the significant temporal features of passenger flow for short-term URT passenger flow prediction from global and local perspectives.

The attention mechanism, also called scaled dot-product attention in the origin paper, can be regarded as a mapping function. In the conventional attention mechanism of the Transformer, linear projection is used to generate the trainable Q, K, and V for discrete inputs. However, linear projection cannot capture the local evolution features inherent in continuous traffic flow data, which may result in the wrong allocation of temporal attention weights. Since the convolution operation can effectively capture the local features, we apply convolution operation instead of linear projection to generate the learnable Q and K with local evolution features, which can calculate more accurate attention scores. Note that the convolution operation here refers to causal convolution in order to avoid future information being learned.

Let \( Q \in R^{s \times d_q} \) denotes the task-related query vector, \( K \in R^{s \times d_k} \) denotes a key vector, and \( V \in R^{s \times d_v} \) denotes a value vector \( V \in R^{s \times d_v} \). They are obtained from the passenger feature matrix \( Z_{alt} \in R^{s \times d_v} \) by causal convolution. In our model, Q, K, and V are the same, and the input to produce them is a sentence, but a time series matrix \( Z_{alt} \). This latent subspace process can be formulated as follows:

\[
Q = \sigma(W_q \ast Z_{alt} + b_q) \quad (14)
\]

\[
K = \sigma(W_k \ast Z_{alt} + b_k) \quad (15)
\]

\[
V = \sigma(W_v \cdot Z_{alt} + b_v) \quad (16)
\]

Where \( \sigma \) denotes the activate function, "\( \ast \)" denotes the causal convolution operation, "\( \cdot \)" denotes the linear projection, and \( W_q, W_k, W_v \) denote the learnable weight for \( Q, K, V \) respectively. After getting \( Q, K, V \), we can calculate the temporal dependencies \( Z \) by dot-product as:

\[
Attention(Q, K, V) = Z = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (17)
\]

where \( \text{Softmax}(\cdot) \) is the activation function to make the nonlinear transformation of the input so that the temporal dependencies \( Z \) is between \([0, 1]\), and \( d_k \) denotes the dimension of the \( Q, K \) matrix, which is used to avoid the gradient too small to backpropagate.

To capture multiple temporal features of passenger flow for increasing the prediction accuracy, the multi-head attention mechanism consisting of multiple self-attention is utilized in this work. The process of the multi-head attention mechanism is shown in Figure 4. Firstly, the passenger flow matrix \( Z_{alt} \) is input to \( h \) different self-attention respectively to calculate the \( h \) temporal feature matrices. Then the \( h \) temporal feature matrices are concatenated together and input into a fully connected layer to obtain the final temporal feature matrix \( Z_h \).

D. Feature Fusion Block

Social media have proven to be a reliable data source to assist in the passenger flow prediction task. It can provide an economical and effective method to obtain traffic-related data. Therefore, an increasing number of studies integrated social media into their models to improve prediction performance.

In this section, we develop a Feature Fusion Block to extract the temporal features from social media to reflect the trend of passenger flow over time. Its calculation process is shown in Figure 5. Specifically, we first use learned embeddings to convert the social media matrix \( S^{t-1} \) to feature matrix \( F^{t-1} \) of dimension \( d_{model} \) to refine temporal features. Then, we apply 2D CNN with pooling operations along the temporal axis for feature matrix \( F^{t-1} \) to capture the trend of passenger flow over time from social media. This feature extraction operation can be formulated as follows.

\[
\hat{F}^{t-1} = \text{Pooling}[f^{7 \times 7}(F^{t-1})] \otimes F^{t-1} \quad (18)
\]

where Pooling denotes the average pooling operation here, \( f^{7 \times 7} \) represents a convolution operation with \( 7 \times 7 \) filter, and \( \otimes \) denotes the element-wise multiplication. Then, a fully connected layer is used to project the processed feature matrix \( F^{t-1} \) to the sample space. Finally, we fuse the processed feature matrix \( F^{t-1} \) and the output \( F^{t-1} \) of residual block by concatenating operation to obtain the feature fusion matrix.
V. EVALUATION

In this section, we will verify the feasibility of the STAFN with two real-world datasets. We first introduce the dataset used in our study, then our model configurations and the evaluation metrics are described. We also choose several conventional models as benchmark models to compare the prediction performance. Finally, the experimental results are analyzed from several perspectives.

A. Dataset

Our study is based on a real-world city URT network from 2019 to 2021 in China. We focus on the subway AFC data, which contains passenger flow data for five weeks from 6:00 a.m. to 11:00 p.m. around New Year’s Day in 2019, 2020, and 2021 as shown in Table I. For adapting our proposed model, we divide the passenger flow data by 10-minute intervals and number the stations uniquely as shown in Table II. To better learn the holiday characteristics of passenger flow on New Year’s Day, we utilize the two consecutive New Year’s Day passenger flow for training and modeling. Take the passenger flow data around New Year’s Day in 2019 and 2020 as an example. The passenger flow data in the first nine weeks, which contains the passenger flow around New Year’s Day in 2019, are used to train and validate the model. The rest data in the last week around New Year’s Day in 2020 are used to test the model. To ensure the accuracy of prediction, our model only considers the consistent stations from different years. The details of the two datasets utilized in our paper are shown in Table III.

We crawled the social media data with the keywords “New Year holiday” and “Nanning” from social media (for example Sina microblogs and Twitter) during a specific period. The specific period is consistent with the time period of URT passenger flow data. For easy calculation, we assume the influence of social media on the passenger flow of all URT stations are the same. Because the sample size of social media data we crawled is too small, it is not appropriate to extract the volumes at specific time granularity, resulting in a sparse matrix. Therefore, we assume that the daily social media volumes have a consistent impact on passenger flow at each time interval of the day.

Here, we briefly analyze the time series data. Figure 6 shows the comparison between the inflow data and the social media volumes at Pengfei Road station during 2021 New Year’s Day. There are obvious peak characteristics of the Nanning URT passenger flow on weekdays, and the holiday passenger flow is significantly more than the weekday passenger flow. With the arrival of New Year’s Day, the social media volume has gradually increased, which is consistent with the distribution of the passenger flow during New Year’s Day, indicating that there is a potential correlation between URT passenger flow and related social media data.

| Card number | Tap-in time | Tap-in station | Tap-out time | Tap-out station |
|-------------|-------------|----------------|--------------|-----------------|
| 3099**      | 2018/12/31  | Xiuxiang       | 2018/12/31   | Chaoyang Square |
| **          | 06:26:18    |                | 06:40:35     |                 |
| 3093*       | 2018/12/31  | Xinmin         | 2018/12/31   | Macun Square    |
| ***         | 19:45:17    | Road           | 20:00:30     |                 |
| 3150*       | 2018/12/31  | Baicanglin     | 2018/12/31   | Jinhu Square    |
| ***         | 07:26:33    |                | 07:43:55     |                 |
| ...         | ...         | ...            | ...          | ...             |
| ...         | ...         | ...            | ...          | ...             |
| 76          | 1           | 2              | 34           | 5               |

# TABLE II INBOUND PASSENGER FLOW STATISTICS OF 2021

| Station Index | 06:00-06:10 | 06:10-06:20 | 06:20-06:30 | ... | 22:50-23:00 |
|---------------|-------------|-------------|-------------|-----|-------------|
| 1             | 0           | 12          | 29          | ... | 0           |
| 2             | 0           | 7           | 16          | 1   | 2           |
| 3             | 2           | 10          | 18          | 1   | ...         |
| ...           | ...         | ...         | ...         |     | ...         |
| 76            | 1           | 2           | 34          | 5   | ...         |

# TABLE III DATA DESCRIPTION

| Description | 2019, 2020 New Year’s Day | 2020, 2021 New Year’s Day |
|-------------|---------------------------|---------------------------|
| Date        | December 3, 2018 to January 6, 2019 | December 2, 2019 to January 5, 2020 |
| Time in a day | 06:00 to 23:00            | 06:00 to 23:00            |
| Line number  | 2                         | 3                         |
| Station number | 41                       | 61                        |
| Time interval | 10                       | 10                        |
| Data record  | 55 million                | 73 million                |
| Week number  | 5                         | 5                         |
| Day number   | 35                        | 35                        |
| Time slice in one day | 102                   | 102                       |
| Total time slice | 3570                  | 3570                      |
B. Model configurations and evaluation metrics

In this paper, all models are implemented with PyTorch on a desktop computer with Intel® Core™ i9-10900X CPU, 64 GB memory, and an NVIDIA GeForce RTX3050 GPU.

Hyperparameters: The same parameters of our STAFN model are applied for both of the two datasets to evaluate its performance. As mentioned above, our model consists of three components. In MGATN, the input dimension and output dimension are consistent, both are (batch size, number of stations, time steps). The Conv-Attention Block consists of three identical conv-attention layers, and each conv-attention layer mainly consists of three-heads attention mechanisms. And the parameters of convolution filters to generate Q and K are kernel size =3, stride = 1, and padding = 1, respectively. In Feature Fusion Block, we set \( d_{model} \) as 24, the parameters of 2D convolution filter are kernel size = 7, stride=1, and padding = 3. The fully connected network consists of two hidden layers and one output layer, the unit numbers of the two hidden layers are [2048, 1024] respectively. And we adopt the RELU function as the activation function for the fully connected network. The batch size is 64. The optimizer is Adam with a learning rate of 0.0001. In addition, we choose 10 minutes as time interval to extract the passenger flow matrix and we use historical 12-time steps after fine-tuning to predict passenger flow \( Y_t \) at the next time step \( t \).

Preprocessing: Before training, all data are normalized to range (0,1) with Min-Max normalization scalers. The results evaluation is conducted after the predicted results are re-scaled to their original scale range.

Evaluation Metrics: For evaluating the prediction performance of the implemented model, we use Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Weighted Mean Absolute Percentage Error (WMAPE) as performance measures. The mean-squared error (MSE) is used as the loss function. When the predicted value is the same as the real value, these evaluation metrics are equal to 0. The closer the values of the three indicators are to 0, the better the prediction accuracy of the model is. We choose the model considering the best performance overall models with different parameters. And the formula of performance measures is as follows:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}
\]

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \hat{Y}_i|
\]

\[
WMAPE = \frac{\sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{Y_i}}{\sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{\hat{Y}_i}} \times 100\%
\]

\[
Loss = MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]

where, \( Y_i \) is the actual passenger flow at time \( i \), \( \hat{Y}_i \) is the predicted passenger flow at time \( i \) and \( n \) is the number of stations.

C. Benchmark models

In this section, we compare the prediction performance of our proposed STAFN model with several benchmark models on Nanning subway passenger flow data for 2020 and 2021 New Year’s Day. The details of the benchmark models are described as follows:

**Autoregressive Integrated Moving Average (ARIMA):**
Autoregressive Integrated Moving Average is well known as a conventional time-series prediction method and is widely used in passenger flow prediction. The three parameters in ARIMA, namely the lag order, the degree of difference, and the order of the moving average, are set as 2, 1, and 0, respectively after fine-tuning.

**Back Propagation Neural Network (BPNN):** As one of the most conventional neural network models, BPNN has been proven to be applicable in predicting passenger flow. We apply a BPNN model consisting of two fully connected layers with 512 neural units in each layer. The optimizer is Adam with a learning rate of 0.0001. We use passenger flow data of full network stations to train this model. The inputs are the inflow of the last 12 time steps. The output is the inflow of the next time step.

**Convolutional Neural Network (CNN):** Well known for the great performance of capturing spatial correlation, CNN has already been applied to time-series data processing. We apply a general 2D CNN model with one CNN layer and two fully connected layers. The parameters of the CNN layer are out_channels = 8, kernel_size =3, stride = 1 and padding = 1. The optimizer is Adam with a learning rate of 0.0001. The inputs and outputs are the same as BPNN.

**Long Short-Term Memory Neural Network (LSTM):** As a deep learning model for processing sequence data, LSTM is widely used in passenger flow prediction. Specifically, we establish an LSTM model with two hidden layers and two fully connected layers. Each LSTM layer consists of 128 neural units. The optimizer is Adam with a learning rate of 0.0001. The inputs and outputs of LSTM are the same as BPNN.

**ST-GCN:** Proposed by Yan et al.(2018)ST-GCN can automatically learn the spatial and temporal features from passenger flow data. We adopt three branches of spatial-temporal graph convolution units. The other parameters are similar to Yan et al.

**GCN-CNN:** A deep learning architecture composed of GCN and 3D CNN. The parameters are the same as Zhang et al.(2020)

**Graph WaveNet (GWN):** Developed by Wu et al., GWN adopts an adaptive dependency matrix to capture dynamic spatial correlation. It also utilizes TCN block to model the temporal features. We implement GWN according to its code.

**Diffusion Convolutional Recurrent Neural Network (DCRNN):** Proposed by Li et al., DCRNN regards traffic flow as a diffusion process on a directed graph, which is
beneficial to model the spatial dependencies. We implement DCRNN based on its code.

**ST-ResNet:** Proposed by Zhang et al. (2017) it has been proved to outperform well-known methods in citywide crowd flow prediction. Here, we only adopt three branches of residual convolutional units and do not consider external factors. The other parameters are the same as his paper.

**ConvLSTM:** Proposed by Shi et al. (2015), ConvLSTM can fully capture the temporal and spatial characteristics of passenger flow. We set up a ConvLSTM model with three hidden layers and two fully connected layers. Other settings are the same as his paper.

**Transformer:** Proposed by Vaswani, Transformer is gradually applied to time series processing because of its multi-head attention mechanism. We apply the encoder-decoder framework with a total of three layers. In each layer, we adopt three multi-head attention, and the values of \( d_k \) and \( d_v \) are all set to eight. We implement Transformer based on its code.

**D. Experiment results**

1) **Network-wide prediction performance**

Table IV and Figure 7 show the prediction performance of the STAFN model and other benchmark models in two real-world datasets. As shown in Table IV, the deep learning models significantly outperform the mathematical statistics-based model, namely ARIMA, which performs the worst no matter in which dataset, with the highest RMSE of 63.291 and 64.452, and the highest MAE of 38.929 and 34.678, respectively. The reason is that ARIMA cannot capture complex nonlinear relationships of passenger flow, and it can only capture limited temporal correlations. Thus, it is not suitable for ARIMA to predict the complex dynamic passenger flow during holidays.

Furthermore, we compare our model with other deep learning-based prediction models. Among these deep learning-based models, BPNN could only capture limited nonlinear characteristics, so its prediction performance is only better than that of ARIMA, while CNN can capture more spatial correlations and LSTM can capture more temporal correlations so they perform better than BPNN. Complex deep learning architectures like ST-GCN, GCN-CNN, GWN, DCRNN, ST-ResNet, and ConvLSTM, which consider both spatial and temporal dependencies, have realized prediction accuracy improvements over the models mentioned above. However, as these models are proposed to predict the passenger flow on general weekdays, they are not suitable to be applied to predict the passenger flow on holidays. The attention-based model Transformer achieved the second-best performance. This is because the attention mechanism can capture the temporal dependencies effectively. However, the conventional attention mechanism ignores the local temporal features, resulting in some mismatches of attention weight. Hence, we especially propose a deep learning-based model to address the limitations mentioned above.

In STAFN, we develop a novel multi-graph attention network (MGATN) to capture the dynamic multiple spatial dependencies of passenger flow. The adjacent graph \( G_a \), functional similarity graph \( G_s \), and OD correlation graph \( G_c \) are applied into MGATN to model the multiple spatial correlations of passenger flow dynamically. As we have introduced, the conv-attention block has a great potential to capture temporal features from global and local perspectives. Therefore, we utilize the convolution operation instead of the linear projection to produce the learnable Q, K with local evolution features for attention weight calculation.

In addition to improving in model structure, we also extend the data sources to improve the prediction accuracy. As existing studies have shown (Ni and He et al., 2017; Xue and Liu et al., 2022), there exists a moderate positive correlation between event passenger flow and related social media data volumes. In our model, we also develop a novel feature fusion block to consider the impact of related social media for passenger flow which can explore the evolution features of passenger flow on holidays, thus improving the accuracy of the passenger flow prediction on holidays.

As we expect, the proposed model STAFN achieves the best prediction performance compared with the benchmark models in most cases.

![Figure 7. Comparison of prediction performances for different model](image-url)
2) **Prediction performance of individual stations**

During holidays, not all subway station passengers have obvious holiday characteristics. The passenger flow of those stations adjoining the business district may have apparent holiday characteristics, while the passenger flow of those stations that undertake daily commuting or connect urban and suburban areas do not. In this paper, we choose three stations with different passenger flow patterns to show the prediction performance of the STAFN. The first station is TingHong Road station, which is adjacent to the main business district, and many citizens visit here on holidays. The second station is Guangxi University station, a typical commuter station with many passengers living nearby. The last station is Nanning Railway station, which is a large transfer hub that can achieve the transfer among various modes of transportation. The prediction results of these three stations during the New Year’s holiday are shown in Figure 8. Below are the results analysis.

The prediction result of TingHong Road station is shown in Figure 8(a). It can be seen that the passenger flow during New Year’s holiday present apparent holiday characteristics: the peak passenger flow is more obvious and much larger than usual. In this case, our proposed model can fully capture the holiday characteristics of passenger flow, with the prediction results closely aligned with the actual values.

The prediction result of Guangxi University station is shown in Figure 8(b). It can be seen that the passenger flow in Guangxi University station has significant commuting characteristics, including obvious morning and evening peak characteristics. During the New Year’s holiday, the number of commuters has dropped significantly. As the STAFN can capture these commuting passenger flow features, the prediction performance is favorable no matter on the peak period, holidays, or weekdays.

The prediction result of Nanning railway station is shown in Figure 8(c). As a transfer hub, the peak-hour distribution of passenger flow is inconsistent with the usual. There is no obvious rush hour in the morning, instead, there are two peak periods in the afternoon and evening. During New Year’s holiday, the increase on the number of tourists from other places has also led to an increase of the subway passenger flow. It can be seen from the Figure 8(c) that the predicted value of our model is close to the actual value, which indicates that the STAFN can capture the spatiotemporal characteristics of various types of passenger flow.

In summary, the STAFN model has strong robustness and it can achieve accurate short-term passenger flow prediction not only for the whole urban rail transit network but also for different types of stations.

![Figure 8](image_url)

**Figure 8** Comparison of actual and predictive values of the three selected stations in Nanning Subway New Year 2020 and 2021: (a) Tinghong Road station; (b) Guangxi University station; (c) Nanning Railway station

3) **Prediction performance in different time intervals**

To further study the prediction performance of the STAFN in different time intervals of a day, we calculate the average loss at each time interval from 6:00 to 23:00 for both two datasets. The prediction performance between our proposed model and other benchmark models (except for ARIMA and BPNN,
because their metrics are too large to display) at different time intervals is described in Figure 9. Below are several conclusions.

Firstly, we compare the prediction performance of the time intervals and the overall prediction performance. As Figure 7 and Figure 9 show, the overall prediction performance of all the models has the same pattern as the performance of different time intervals. Our model STAFN outperforms the benchmark models over most of the time intervals of a day with the lowest average evaluation metrics, which is consistent with the overall prediction performance of STAFN. Results illustrate that STAFN has stable predictive performance, and it is suitable for network-wide prediction.

Secondly, the prediction performance of the same model in different time intervals of a day will be analyzed. It can be seen from Figure 9 that, there exists a roughly similar trend in the evaluation metrics for all models. Consistent with the morning and evening peak features of passenger flow, the evaluation metrics of different time intervals in a day have an obvious peak period, indicating that the evaluation metrics go up when the passenger flow increases sharply. Compared with other baselines, the evaluation metrics of STAFN model have the slightest morning and evening peak features in time interval-level prediction, which illustrates that the STAFN is strongly stable and robust no matter in peak period or off-peak period.

Then, we analyze the prediction performance of different models in the same time intervals of a day. No matter whether in peak period or off-peak period, ARIMA performs the worst among all models with the highest evaluated metrics value. While the deep learning models outperform the statistics-based model with much lower metrics value. The performance of these deep learning models is roughly similar. Among these models, the STAFN performs the best over most of the time intervals during a day, which means that our model is also suitable for passenger flow prediction in a single time interval. Eventually, the prediction performance of all models on different datasets is discussed. It can be inferred that there exists similar prediction performance of all models for both of the datasets in 2020 New Year’s holiday and 2021 New Year’s holiday, which suggests that the prediction performances of all models are generally consistent in different datasets.

To sum up, our proposed STAFN can achieve favorable prediction performance on holidays in most cases.

4) Prediction performance of different ablation studies
To further explore the influence of social media data and model architecture for prediction accuracy, we conduct detailed ablation studies. Firstly, we remove parts of the STAFN structure and keep other parts unchanged to evaluate their impacts on the prediction performance. Then, we test the prediction performance of the model using different data sources, such as using one year or two consecutive years of passenger flow, with and without using microblog data. Finally, we compare the RMSE, MAE, and WMAPE of the prediction results. The experiment settings can be summarized as follows.

(1) Using the conventional attention mechanism: replace conv-attention with conventional attention mechanism.
(2) Without conv-attention block: remove conv-attention block from STAFN;
(3) Using conventional GCN: replace MGATN with the conventional GCN;
(4) Without MGATN structure: remove MAGTN structure from STAFN;
(5) Without Multiple Graphs: only adopt the adjacent graph to model the spatial dependences;
(6) Without ResNet Block: remove ResNet Block from STAFN;
(7) Without Feature Fusion Block: do not fuse the microblog data into STAFN, only use the passenger flow data;
(8) Using one-year data: use only current year data rather than two consecutive years data.

Table V and Figure 10 illustrate the experimental results of different ablation studies. It can be seen from the results that the STAFN model outperforms its variant models with the lowest RMSE of 26.606 and 29.912, the lowest MAE of 16.306 and 15.826, and the lowest WMAPE of 0.124 and 0.158 in both the two datasets, respectively, indicating that the proposed model is the best configuration among variants, and all components of STAFN are indispensable.

Specifically, lacking ResNet block in STAFN results in the second worst prediction performance, which shows that the ResNet block is important for our model. Because it can contribute to easily optimizing the model and enhancing representation ability, thus significantly improving prediction accuracy.

In terms of attention mechanism, when we do not leverage the attention mechanism in the STAFN, the prediction performance has decreased significantly. This is because when we remove the attention mechanism, our model can hardly capture the complex temporal dependencies of passenger flow. When we replace conv-attention with the conventional attention mechanism in STAFN, the prediction performance has also decreased, which indicated that the conv-attention block can enhance the ability to capture the temporal dependencies from global and local perspectives, thus helping to improve prediction accuracy.

![Figure 9 Comparison of the performance in different time interval of a day.](image-url)
TABLE V COMPARISON OF PREDICTION PERFORMANCE UNDER DIFFERENT CONDITIONS. THE BEST RESULTS ARE EMPHASIZED IN BOLD FONT.

| Model                        | 2020 New Year |         | 2021 New Year |         |
|------------------------------|---------------|---------|---------------|---------|
|                              | RMSE          | MAE     | WMAPE         | RMSE    | MAE     | WMAPE |
| STAFN-noResNet               | 46.976        | 27.549  | 0.208         | 56.449  | 26.980  | 0.271 |
| STAFN-noAttention            | 30.873        | 18.869  | 0.143         | 32.544  | 17.912  | 0.179 |
| STAFN-noGNNs                 | 31.232        | 19.256  | 0.145         | 31.653  | 18.221  | 0.181 |
| STAFN-Conventional Attention | 29.728        | 17.448  | 0.132         | 29.907  | 16.031  | 0.161 |
| STAFN-Conventional GCN       | 30.423        | 18.231  | 0.137         | 30.699  | 16.564  | 0.165 |
| STAFN-single graph           | 30.281        | 18.024  | 0.137         | 30.308  | 16.287  | 0.163 |
| STAFN-noFusion               | 29.293        | 17.155  | 0.130         | 30.922  | 16.602  | 0.164 |
| STAFN-one year data          | 83.119        | 54.799  | 0.540         | 56.951  | 34.113  | 0.40  |
| STAFN                        | **28.880**    | **17.040** | **0.128**    | **28.756** | **15.953** | **0.159** |

In terms of GNNs, when we remove the GNN block, we can find that the prediction accuracy of STAFN becomes much worse because of the poor ability to model the spatial dependencies for STAFN without GNNs block. When we utilize conventional GCN instead of MGATN in STAFN, the prediction performance is worse than the STAFN with MGATN block, showing that the novel MGATN can model the multiple spatial dependencies dynamically according to the input, which can enhance the prediction ability of STAFN. When we only apply a single graph to model spatial dependencies, our model can only capture limited spatial features, resulting in poor prediction results.

We also compare the prediction performance of using one year’s or two consecutive years’ passenger flow data. It can be seen from Table V and Figure 10 that when using one year’s passenger flow, the prediction accuracy of the STAFN is the worst among all variants, showing that using two consecutive years’ data to conduct the prediction can better capture the holiday features of passenger flow, thus solving the problem of insufficient prediction accuracy caused by the limited sample size of holiday passenger flow. We also compare the prediction performance of STAFN with and without feature fusion block. When we adopt the feature fusion block into STAFN to explore the impact of related social media on holiday passenger flow, STAFN can better learn the evolution trend of holiday passenger flow over time. It can be shown that it is beneficial to fuse related social media to predict holiday passenger flow.

All of the above illustrates that our model architecture can fully capture the complex spatial-temporal dependencies of holiday passenger flow, and the related social media volumes can learn the temporal trend of holiday passenger flow well, which ensures the favorable prediction performance of holiday passenger flow.

VI. CONCLUSION

Predicting short-term passenger flow on holidays for the URT system is a significantly challenging task for traffic management because of its suddenness and irregularity. In our study, we develop a deep-learning architecture called STAFN to conduct the short-term passenger flow on holidays. The main conclusions are summarized as follows.

1. The proposed STAFN has significant advantages to capture spatial-temporal dependencies of passenger flow, especially on holidays and topological information of the subway network.

2. The STAFN utilizing the microblog volumes outperforms other benchmark models and achieves favorable prediction accuracy. The improvements compared with the best (existing) models are RMSE of 4.02%, MAE of 5.98%, and WMAPE of 5.35%, respectively. This indicates that social media data can be regarded as an effective data source to improve passenger flow prediction.

3. The results tested on two real-world datasets reveal that the STAFN performs well under different ablation studies, showing the favorable robustness and the great potential to be applied in the real world.

However, there are several limitations to our study. For example, we only use a single feature, social media volumes, when capturing the factors affecting passenger flow. In the future, we will consider multi-features, such as emotions of social media users, weather conditions, and date attributes, which may improve the accuracy of the prediction. Besides, whether the proposed model can be applied to other scenarios, such as speed prediction is also worth studying in future work.

Figure 10 Performance comparison of different condition
Zheng, W. & Lee, D. (2006), "Short-Term Freeway Traffic Flow Prediction Bayesian Combined Neural Network Approach", Journal of transportation engineering, Vol. 132 No.2006, pp. 114-121.

Shuxin Zhang was born in Guangdong Province, China. He is a master in Beijing Jiaotong University. His research interests include machine learning, deep learning, traffic data mining and applications, and dynamic traffic modeling and management.

Jinlei Zhang was born in Hebei Province, China. He received a Ph.D. degree from Beijing Jiaotong University, China. He is currently an assistant professor at Beijing Jiaotong University. His research interests include machine learning, deep learning, traffic data mining and applications, and dynamic traffic modeling and management.

Lixing Yang received the B.S. and M.S. degrees from the Department of Mathematics, Hebei University, Baoding, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Department of Mathematical Sciences, Tsinghua University, Beijing, China, in 2005. Since 2005, he has been with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, where he is currently a Professor. He is the author or co-author of more than 80 papers published in national conferences, international conferences, and premier journals. His current research interests include stochastic programming, fuzzy programming, intelligent systems, and applications in transportation problems and rail traffic control systems.

Jiateng Yin received his Ph.D. degree from the School of Electronic information engineering, Beijing Jiaotong University in 2018. He is also a co-doctoral student of the University of Wisconsin-Madison. He is currently a Associate Professor with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. His research interests include artificial intelligence, traffic intelligent control and optimization, rail transit system safety and control engineering.

Ziyou Gao received the Ph.D. degree from the Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, China, in 1994. He is currently a Professor with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University. He also serves as the President of the Society of Management Science and Engineering of China, the Vice-President of the Systems Engineering Society of China, and the Vice-President of the Chinese Society of Optimization, Overall Planning and Economic Mathematics. In 2003, he was elected as a foreign member of the Russian Academy of Natural Sciences. His research interests include management and optimization of urban rail transit, complexity of transportation network, and transport modeling and planning.