Metro Ridership Forecasting using Inter-Station-Aware Transformer Networks

Khaled Saleh[1], Adriana-Simona Mihaita[2] and Yuming Ou[2]
[1] School of Information and Physical Sciences, University of Newcastle, Australia
[2] Faculty of Engineering and IT, University of Technology Sydney, Australia
Email: khaled.saleh@newcastle.edu.au

***Abstract*— In recent years, the issue of predicting metro ridership has gained traction within the intelligent transportation systems community, due to its potential advantages for the metro network system such as improving the service quality and making informed decisions about infrastructure investments. When it comes to metro station-level ridership forecasting, in the literature this is often tackled by using recurrent neural network (RNN)-based approaches. While RNNs have shown promising results in providing station-level metro ridership predictions over short-term prediction horizons, they are still challenged when it comes to long-term prediction horizons. Thus, in this work, we are introducing a novel approach, the Inter-Station-Aware Transformer Networks framework, for efficient and scalable station-level metro ridership forecasting over both short and long-term prediction horizons. Our proposed approach models and fuses both the temporal historical ridership data and the metro network topology using an encoder-decoder framework based on the transformer network architecture. The proposed approach has been evaluated on two publicly available datasets and compared against a number of baseline approaches. We achieved superior results when it comes to longer-prediction horizons when compared with state-of-the-art methods from the literature, while we proved it is also three times more efficient in terms of the number of model parameters required.**

I. Introduction

Metro systems are critical transportation infrastructure in many urban areas, providing millions of people with reliable and efficient mobility options. Furthermore, metro systems can alleviate traffic congestion, enhance capacity on heavily used routes, and decrease the emission of local pollutants and greenhouse gases [1]. That being said and given the immense number of ridership happening on the metro system networks, the risk of citywide congestion still exists if the service and management of the network are inefficient. Thus, understanding and accurately predicting metro ridership (e.g., passengers' inflow and outflow) is considered one of the key enablers for efficient operational planning, resource allocation, and urban transportation management [2]. Metro ridership forecasting on the station level particularly presents unique challenges due to the complexity and dynamics of urban transportation systems. Despite the challenges, accurate and reliable ridership forecasting on the station level has significant benefits. It allows metro operators to optimize resources, such as scheduling of trains, allocation of staff, and maintenance planning, resulting in improved service quality and cost-effective operations. Additionally, it enables policymakers to make informed decisions about station-specific infrastructure investments, such as platform expansions, station renovations, and accessibility enhancements, to meet the changing needs of passengers. Over the past few years, various approaches have been proposed for metro ridership forecasting which is mostly inspired by the work on traffic state estimation problems (i.e. speed, flow or demand). Those approaches range from traditional statistical methods [3], [4] and machine learning methods [5]–[7] to hybrid models that combine different methodologies [8], [9]. These approaches utilize various data sources, including historical ridership data, socio-demographic data, weather data, and other relevant factors, to develop accurate and reliable forecasting models. However, there is still a need for further research and development to address the unique challenges associated with station-level ridership forecasting. One of the key challenges is how to achieve an efficient representation and modelling of spatial and temporal dynamics of the metro system networks that are scalable across different network topologies. Recently, machine-learning-based techniques especially those based on Graph Convolution Networks (GCN) [7], [10] have shown promising results for both road and rail-based ridership forecasting problems. However, given the spatio-temporal nature of the problem, GCN is commonly paired with another type of machine learning model (such as recurrent neural networks [11]) that can handle the temporal aspect of the problem because GCN can only handle the spatial aspect. Consequently, this over-complicates the final forecasting model and makes it inefficient while handling large-scale ridership data.

Thus, in this work, we are proposing a novel single unified data-driven approach based on transformer networks [12] that can efficiently handle the inherent spatio-temporal nature of the station-level metro ridership forecasting problem. The proposed approach will model and fuse both the temporal historical ridership data and the metro network topology using an encoder-decoder framework based on the transformer network architecture. The remaining sections of the paper are structured as follows. Section II provides a brief overview of the relevant literature around the metro ridership forecasting problem. Section III outlines the proposed approach and methodology. Section IV presents the experimental results and performance evaluation of the proposed approach. Finally, Section V concludes the paper.

Manuscript 470 submitted to 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). Received May 19, 2023.

II. Related Work

In the literature, the work done on the metro ridership forecasting problem can be categorised into three main categories, namely traditional statistical methods, machine learning-based methods and hybrid model methods. One of the early statistical methods work was presented by Li et al. [3], where a multi-scale radial basis function (MSRBF) network for predicting short-term subway passenger flow during special events using smart card data was proposed. Similarly, in [4] authors proposed statistical models for modelling passenger flows based on smart card's tap-on and tap-off times. Then, they used Bayesian inference to estimate parameters in these models. On the other hand, machine learning-based (ML) methods utilise a purely data-driven approach where, given relatively large-scale data sets, the ML models are trained/optimised using this available information. Gong et al. [5], proposed a customized online non-negative matrix factorization (ONMF) method for short-term prediction of crowd flow distribution across the entire Sydney trains network in Australia. Given that their method was optimised using querying average historic ridership data of the day where the forecasting is performed, they have developed two separate models that can provide ridership forecasting during normal weekday operations and another one during the peak/rush hours. On the contrary, in [6], they have proposed a deep learning framework for only short-term (next 15 minutes) metro network-wide passenger flow prediction based on a long short-term memory (LSTM) model. In their model, they only took into account historical ridership data and time data without any spatial information about the metro network topology itself. Similarly, in [7], another deep learning-based method based on LSTM architecture was proposed. However, unlike [6] they have considered the physical metro network topology by complementing the LSTM model with another machine learning model based on the GCN architecture which has improved the overall performance of their model for both short/long-term prediction horizons. Other recent studies have also looked into coupling train patronage prediction using CNNs-LSTMs in conjunction with large Digital Twin Models powered by mobile data as well as train real-time movements (see [13]), or even integrating the patronage movement into a dynamic passenger assignment (see [14]) or for studying the impact of train disruptions on the train network patronage and scheduling (see [15], [16]).

Rather than the purely statistical or machine learning-based methods, in the hybrid model methods, a combination of the two methods is performed. For example, in [8], a framework that combines both Wavelet and SVM models for short-time passenger flow prediction in Beijing's subway system. Similarly, in [17], another hybrid one-step subway short-term ridership forecasting method was presented. Their approach accounts for ridership dynamics and uncertainties, using the autoregressive integrated moving average (ARIMA) model, and the non-linear generalized autoregressive conditional heteroskedasticity (GARCH) family model. While all these methods can provide relatively good results when it comes to metro ridership forecasting, however, as we have highlighted they are either focus on short/long-term prediction horizons and for those approaches how can perform short and long-term predictions they are not efficient enough to be deployable and scalable across different metro network systems.

III. Proposed Methodology

In this section, a formal definition of station-level metro ridership forecasting is first provided. Then, a brief overview of the input metro network topology representation for our proposed methodology will be discussed. Lastly, the full details of our proposed novel topology-aware transformer networks-based model will be presented.

A. Problem Statement

In our formulation for the station-level metro ridership forecasting problem, on a high level, the problem is cast as a time-series forecasting problem where, given an input sequence (with the length of n) of historical metro ridership data $\left(D^N_{t-n+1}, D^N_{t-n+2}, \ldots, D^N_t\right)$ along with the metro network topology graph G, the goal is to predict the sequence (with the length of m) of future ridership data $\left(\hat{D}^N_{t+1}, \hat{D}^N_{t+2}, \ldots, \hat{D}^N_{t+m}\right)$. Based on this formulation, D^N_t represents the ridership data for the total number of stations N at a given time-step t, and this ridership data is composed of two real values ($D \in \mathbb{R}^2$) which correspond to the inflow/outflow count of passengers for each station.

B. Metro Network Representation Graphs

One of the most recent and commonly utilised representation techniques for network topologies of both road and rail networks is the graph representation, which was shown to be effective in enhancing the modelling capabilities of many traffic state estimation tasks [7], [18]–[20]. In this representation, the network is organised as a set of nodes interconnected by using a number of edges and each edge has a corresponding weight to it. There are a number of ways to build this graph representation, and one of the most simple and straightforward methods is to directly build the graph based on the geographical and physical topology of the studied network system; however, this simple strategy was proven to be sub-optimal when it comes to capturing the inter-station flow patterns in a metro system network [7]. Thus, in this work, we will be adopting a similar approach to [7] for the metro network representation graphs. In this approach, the metro network is represented by the following three different types of graphs as detailed below.

1) **Physical Topology Graph:** The first type represents the actual physical topology of the network which focuses on capturing the local spatial dependency between nearby stations. Accordingly, the physical topology graph can be represented by calculating the physical edge weight matrix $M_p(a,b)$ between each pair (a, b) of nodes/stations of the total N stations that exist in the entire metro network as follows:

Manuscript 470 submitted to 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). Received May 19, 2023.

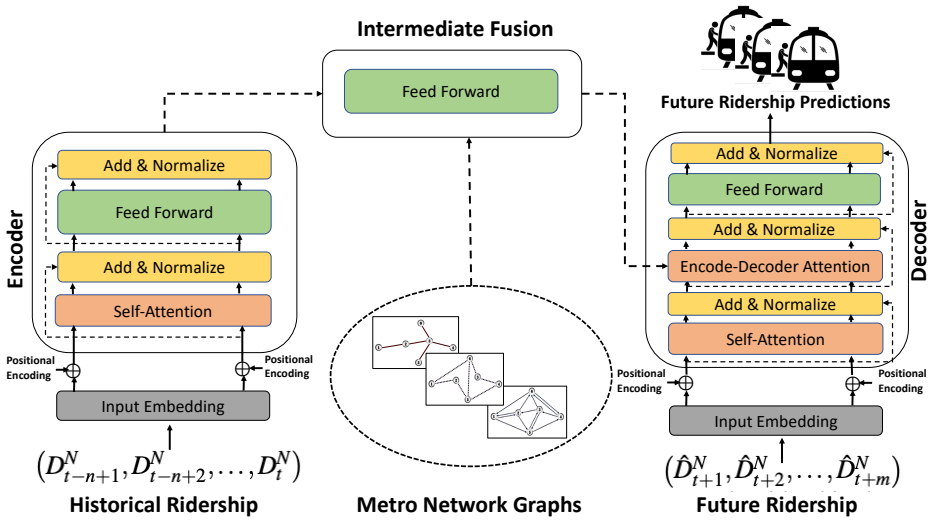

Fig. 1. Our proposed Inter-Station-Aware Transformer Networks framework. The framework consists of three main components, namely encoder, decoder and intermediate fusion.

$$M_p(a,b) = \frac{E(a,b)}{\sum_{k=1}^{N} E(a,k)} \tag{1}$$

where $E(a,b) = 1$ if a physical edge exits between station (a,b), otherwise $E(a,b) = 0$

2) **Semantic Similarity Graph:** The second type of graph is the similarity graph which is constructed based on prior domain knowledge that can capture the semantic similarity between each pair of stations/nodes regarding their passengers' flow. For example, two stations might not be physically interconnected in the metro network, however, both of them are business hub districts, so, in this case, it would be beneficial to connect them in this newly semantic similarity graph in order to capture the ridership flow patterns across them. Consequently, the semantic similarity graph is represented by calculating the semantic similarity edge weight matrix $M_s(a,b)$ as follows:

$$M_s(a,b) = \frac{F_s(a,b)}{\sum_{k=1}^{N} F_s(a,k)} \tag{2}$$

where $F_s(a,b)$ is the passenger flow similarity score between station pair (a,b) which is computed by taking the warping distance, Dynamic Time Warping (DTW) [21] between all historical passenger flow between stations (a,b).

3) **Correlation Graph:** The third and last type of graph is the correlation graph, constructed to capture any correlations between stations. The rationale for constructing a correlation graph is to capture scenarios where the majority of inflow/outflow patterns are happening between two specific stations, as this will have a major influence on the relationships/interactions between stations beyond their physical interconnection. Accordingly, the correlation graph can be represented by calculating the correlation edge weight matrix $M_c(a,b)$ as follows:

$$M_c(a,b) = \frac{R_c(a,b)}{\sum_{k=1}^{N} R_c(a,k)} \tag{3}$$

where $R_c(a,b)$ is the correlation ratio matrix between a station pair (a,b) which is computed by counting the total number of travels that started from station a and ended at station b, normalised by the total number of travels between each pair of stations in the entire metro network.

C. Inter-Station-Aware Transformer Networks Model

The backbone architecture of our proposed framework for metro ridership forecasting is based on the transformer network architecture [12] which recently has achieved state-of-the-art results in both sequential learning and computer vision tasks [22]–[25]. As can be shown in Fig. 1, the proposed model consists of three main components, namely the encoder, the intermediate fusion and the decoder. The encoder and decoder take as input the historical metro ridership data, $(D^N_{t-n+1}, D^N_{t-n+2}, \ldots, D^N_t)$ and the future metro ridership data $(\hat{D}^N_{t+1}, \hat{D}^N_{t+2}, \ldots, \hat{D}^N_{t+m})$ respectively. In order to transform both the input historical/future metro ridership data into a higher dimensional space d_s, a linear embedding layer exists at the beginning of both the encoder and decoder components. Then, the output from the embedding layer (also known as token embeddings) is concatenated with the output from the positional encoding layer, which provides explicit time-tagging/order to the sequential tokens. The encoding technique used in our models is similar to the original transformer network architecture, which is a set of sine and cosine functions with different frequencies, that enables the model to capture positional information at different scales. The intermediate fusion component of our model is then composed of only one type of layer, the feed-forward. As the name implies, the intermediate fusion layer fuses the encoder output and the three types of graphs, described in Section III-B, together via a concatenation operation. Then, the feed-forward layer applies a fully connected neural network to each position in the fused input sequence independently.

On the other hand, both the encoder and decoder, internally are composed of multiple identical layers. The encoder

Manuscript 470 submitted to 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). Received May 19, 2023.

layers include a Multi-Head Attention layer, which computes self-attention to capture the interactions between different positions in the input sequence, a feed-forward layer, which applies a fully connected neural network to each position in the input sequence independently, and a Residual Connection and a Layer Normalization technique to aid the training process. Similarly, the decoder layers include a) a Self-Attention layer, which computes self-attention and also attends to the output of the encoder, b) a Multi-Head Attention layer with both the encoder output layer and the intermediate fusion output layer, which computes the attention between the decoder input the encoder output and the intermediate fusion output. Also, a feed-forward layer and Residual Connection and Layer Normalization techniques exist in the decoder component, which is the same as in the encoder block. Finally, at the end of the decoder component, the output layer exists which is responsible for auto-regressively producing the future metro ridership predictions.

IV. Experiments

In this section, we begin by introducing the datasets that we used to train and evaluate the effectiveness of our proposed method. Next, we describe the details of our experimental setup, including the evaluation metrics and the baselines from the literature that we compared against. Finally, we analyze and discuss the results of our proposed method on two real datasets from two different metro networks.

A. Datasets

To demonstrate the effectiveness of our proposed approach, we will train and evaluate our Inter-Station-Aware Transformer Networks model on two recent datasets which were made publicly available for bench-marking the metro ridership forecasting tasks, namely SHMetro and HZMetro [7]. The two datasets were collected based on large-scale ridership transactions over two big metro networks in two different cities in China. The first dataset, SHMetro is derived from Shanghai's metro system, featuring 811.8 million transactions from July 1st to September 30th, 2016, with an average of 8.82 million daily passengers. Every transaction record contains the passenger ID, the entry/exit train station, and the timestamps. During this period, 288 metro stations were operational, linked by 958 physical edges. Furthermore, the passenger inflow/outflow was measured in 15-minute intervals starting from 5:15 in the morning to 23:30 in the evening for each station within the network. The first two months and the last three weeks of the ridership data were used for training and testing, while the rest was used for validation.

The HZMetro dataset, on the other hand, was collected between the 1st to 25th of January 2019 from the Hangzhou metro system, which covers 80 metro stations linked by 248 physical edges, with an average of 2.35 million daily passengers. Similar to the SHMetro dataset, the time interval for inflow/outflow recording is 15 minutes from 5:15 in the morning to 11:30 in the evening and also the dataset is split over time into three splits for training, validation and testing.

B. Experiment Setup

For our experiments, we started first by pre-processing the two aforementioned datasets. In the first pre-processing stage, we obtained the three metro network representation graphs by computing their edge weight matrices according to the equations described in Section III-B. Furthermore, and similar to [7], for the SHMetro dataset we chose only the highest k (where k=10) stations that have a high degree of similarity scores or correlation rates for calculating their corresponding semantic similarity and the correlation edge weight matrices. The rationale behind this (as it was shown in [7]) is to minimise the computational expenses involved in SHMetro modelling and align with the standard practice of baseline methods in assessing their performance on the SHMetro dataset. Consequently, we ended up with 2880 edges for both the semantic similarity and the correlation graphs. On the other hand, for the HZMetro dataset, we did not adopt the strategy of selecting top k-stations given the lower number of stations and physical edges in comparison to the SHMetro dataset. As a result, we had a semantic similarity graph with 2502 edges and a correlation graph with 1094 edges for the HZMetro dataset. In the second pre-processing stage, we have constructed the passenger's inflow/outflow for each station across the training/validation/testing splits of both the SHMetro and HZMetro to have the past historical metro ridership data length n of 4 which corresponds to 60min given that the time interval for the ridership recording for both two datasets is 15min. While the future metro ridership data length m is also set to 4 during the training and validation phase only, but during the inference/testing stage, our model can auto-regressively forecast any prediction horizon. Given that the future ridership predictions for each metro station from our Inter-Station-Aware Transformer Networks model are continuous values, our selected objective function for training was the $L2$-loss function. The training process used the Adam optimizer and lasted for 250 epochs. In terms of the model's hyper-parameters, we set the hidden units of the input embedding layer to 512 and incorporated 8 self-attention heads into both the encoder and decoder stages.

C. Evaluation Metrics

In accordance with earlier studies [6], [7], the performance of our proposed approach will be assessed using two different evaluation metrics, namely the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) which are computed as follows:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{D}_i - D_i\right)^2}, \\ \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\hat{D}_i - D_i\right|, \end{aligned} \tag{4}$$

where n is the training samples count, D_i is the ground-truth metro ridership and $\hat{D}_i$ is the predicted metro ridership.

TABLE I

QUANTITATIVE EVALUATION OF THE PROPOSED ISA-TF APPROACH IN COMPARISON TO A NUMBER OF BASELINE APPROACHES OVER THE SHMETRO AND HZMETRO DATASETS [7]. THE REPORTED SCORES ARE SHOWN ACCORDING TO THE EVALUATION METRICS RMSE/MAE. LOWER SCORES ARE BETTER. BEST SCORES ARE IN **BOLD**.

Dataset	SHMetro				HZMetro			
Approach	15 min	30 min	45 min	60 min	15 min	30 min	45 min	60 min
HM	136.97/48.26	136.81/47.88	136.45/47.26	135.72/46.40	64.19/36.37	64.10/36.37	63.92/36.23	63.72/35.99
GB	62.59/32.72	82.32/39.50	113.95/49.14	137.5/57.31	51.50/30.88	61.94/36.48	76.70/44.12	91.21/51.10
MLP	48.71/25.16	51.80/26.15	57.06/27.91	63.33/29.92	46.55/26.57	47.96/27.44	50.66/28.79	54.62/30.52
RNN-GRU	52.04/25.91	54.02/26.39	56.97/27.17	59.91/28.08	45.10/45.10	45.26/25.93	46.13/26.36	47.69/26.98
STG2Seq [26]	47.19/24.98	50.58/50.58	52.68/52.68	56.81/28.22	39.52/39.52	40.72/40.72	40.72/40.72	40.72/40.72
DCRNN [27]	46.02/24.04	49.90/49.90	54.92/26.76	58.83/28.01	40.39/40.39	42.57/25.22	46.26/26.97	49.35/28.47
PVCGN [7]	**44.97**/23.29	47.83/24.16	52.02/25.33	55.27/26.29	**37.76/22.68**	**39.34/23.33**	40.95/24.22	42.61/24.93
ISA-TF (ours)	45.85/**23.05**	**46.51/23.27**	**47.38/23.55**	**48.43/23.87**	38.64/23.53	39.56/24.06	**40.27/24.21**	**41.68/24.91**

D. Results and Discussion

In this section, we evaluate the performance of our proposed Inter-Station-Aware Transformer Networks model (we refer to it as ISA-TF) in Table I according to the aforementioned evaluation metrics. Furthermore, we compare the performance of our ISA-TF model against a number of baseline approaches including the state-of-the-art (SOTA) approach (PVCGN) [7] on both the SHMetro and the HZMetro datasets. The compared baseline approaches are as follows:

- **HM**: Historical Mean is a simple statistical model that takes into account the k-historical ridership data of a given time interval by computing their mean to predict the future ridership data at the same time interval. For the SHMetro, k was set to 4 and 2 for the HZMetro.
- **GB**: Gradient Boosting [28] is an ensemble learning method that combines multiple weak learners (in our case decision trees) to form a strong predictive model. It uses gradient descent optimization to find the optimal weights for combining the weak learners. The hyper-parameters for GB were set as follows: 100 for the number of boosting stages and 4 for the max depth of each estimator.
- **MLP**: Multiple Layer Perceptron is a fully-connected neural network architecture which in our case consists of 2 layers with 256 hidden units in the first layer and 2304 and 640 hidden units respectively in the second layer, based on the dataset in use (SHMetro or HZMetro).
- **RNN-GRU**: Gated Recurrent Unit is a recurrent neural network (RNN) based architecture, and in our case, it consists of two GRU layers with its hidden units set to 256.
- **STG2Seq**: Spatial-temporal Graph to Sequence is a method that was first introduced in [26] for passenger demand forecasting. It is a hierarchical graph convolutional model that is composed of two types of encoders (short/long-term) and an attention mechanism for fusion.
- **DCRNN**: Diffusion Convolutional Recurrent Neural Network is a method that was first introduced in [27] for traffic forecasting tasks which exploits random walks on graphs to capture the spatial dependency and utilises the encoder-decoder model for capturing the temporal dependencies.
- **PVCGN**: Physical-Virtual Collaboration Graph Network is the SOTA approach on both SHMetro and HZMetro that was first introduced in [7]. It utilises the same input as our proposed approach (i.e. historical ridership and the three types of metro network graphs), however, its model consists of two types of architectures, namely graph convolution neural networks and RNN-GRU to capture the spatio-temporal dependency of the input data.

As can be seen from the reported results in Table I, our proposed approach ISA-TF has outperformed the baseline approaches over the SHMetro dataset in terms of RMSE and MAE scores. For the HZMetro dataset, our proposed approach has achieved comparable scores to the SOTA approach, PVCGN with a slight improvement for PVCGN over the short-term prediction horizon (i.e. 15 and 30 min), while our ISA-TF achieved better results over the long-term prediction horizon. Moreover, the proposed approach has consistently achieved resilient scores over the long-term prediction horizons across the two datasets, which further proves the scalability of our proposed approach. More importantly, the results also show a key unique property of our proposed ISA-TF which is that it doesn't suffer from the accumulation of prediction errors over time which is quite prevalent with the rest of the baseline approaches (especially those based on graph and recurrent neural networks such as STG2Seq and PVCGN). As a result, this would make our proposed approach more suitable for long-time prediction horizons which are crucial for robust metro ridership forecasting. In order to further prove the efficiency of our proposed approach, in Fig. 2 we highlight the number of parameters required for both our proposed ISA-TF approach and the PVCGN which was the SOTA approach on both SHMetro and HZMetro. As can be seen from the plot, our ISA-TF approach only required 47.6 million parameters for training, while the PVCGN since is composed of two different types of architectures (graph convolution and recurrent neural

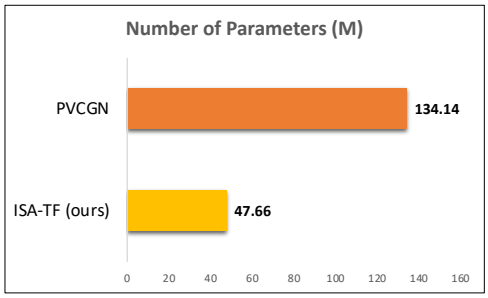

Fig. 2. The number of training parameters comparison between our proposed approach and PVCGN [7]. M refers to million.

networks) requires more than 134 million parameters. This means that our approach is roughly 3 times more efficient when it comes to the number of parameters which plays a critical role in the overall run-time performance of the metro ridership forecasting models.

V. Conclusion

In this work, we proposed a novel single unified data-driven approach, named Inter-Station-Aware Transformer Networks framework for the metro ridership forecasting problem. The performance of our approach has been evaluated on two publicly available datasets for the metro ridership forecasting task. In comparison to the state-of-the-art (SOTA) methods on both datasets, our approach has achieved comparable and sometimes better results over the short-term prediction horizons (i.e. 15 and 30 mins ahead), while it achieved superior results over the long-term prediction horizons (i.e. 45 and 60 mins ahead) in terms of both RMSE and MAE scores. Moreover, the proposed approach is three times more efficient than SOTA approach when it comes to the number of training parameters.

References

[1] D. Lin, J. D. Nelson, M. Beecroft, and J. Cui, "An overview of recent developments in china's metro systems," *Tunnelling and Underground Space Technology*, vol. 111, p. 103783, 2021.

[2] X. Yang, J. Wu, H. Sun, Z. Gao, H. Yin, and Y. Qu, "Performance improvement of energy consumption, passenger time and robustness in metro systems: A multi-objective timetable optimization approach," *Computers & Industrial Engineering*, vol. 137, p. 106076, 2019.

[3] Y. Li, X. Wang, S. Sun, X. Ma, and G. Lu, "Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 306–328, 2017.

[4] C. Li, S. Xiong, X. Sun, and Y. Qin, "Bayesian analysis for metro passenger flows using automated data," *Mathematical Problems in Engineering*, vol. 2022, pp. 1–12, 2022.

[5] Y. Gong, Z. Li, J. Zhang, W. Liu, Y. Zheng, and C. Kirsch, "Network-wide crowd flow prediction of sydney trains via customized online non-negative matrix factorization," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 1243–1252.

[6] S. Hao, D.-H. Lee, and D. Zhao, "Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system," *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 287–300, 2019.

[7] L. Liu, J. Chen, H. Wu, J. Zhen, G. Li, and L. Lin, "Physical-virtual collaboration modeling for intra-and inter-station metro ridership prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3377–3391, 2020.

[8] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-svm short-time passenger flow prediction in beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, 2015.

[9] E. Chen, Z. Ye, C. Wang, and M. Xu, "Subway passenger flow prediction for special events using smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1109–1120, 2019.

[10] S. Fang, Q. Zhang, G. Meng, S. Xiang, and C. Pan, "Gstnet: Global spatial-temporal network for traffic flow prediction." in *IJCAI*, 2019, pp. 2286–2293.

[11] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] Y. Ou, A. Mihaita, A. Ellison, T. Mao, L. Zhang, J. Lozoya Santos, S. Lee, and F. Chen, "Rail digital twin and deep learning for passenger flow prediction using mobile data," 2023.

[14] Y. Ou, A.-S. Mihaita, and F. Chen, "Dynamic train demand estimation and passenger assignment," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–6.

[15] D. Zhao, A.-S. Mihăiță, Y. Ou, S. Shafiei, H. Grzybowska, K. Qin, G. Tan, M. Li, and H. Dia, "Traffic disruption modelling with mode shift in multi-modal networks," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2022, pp. 2428–2435.

[16] Y. Ou, A.-S. Mihăiță, and F. Chen, "Chapter 14 - big data processing and analysis on the impact of covid-19 on public transport delay," in *Data Science for COVID-19*, U. Kose, D. Gupta, V. H. C. de Albuquerque, and A. Khanna, Eds. Academic Press, 2022, pp. 257–278. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323907699000104

[17] C. Ding, J. Duan, Y. Zhang, X. Wu, and G. Yu, "Using an arima-garch modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 4, pp. 1054–1064, 2017.

[18] X. Geng, Y. Li, L. Wang, L. Zhang, Q. Yang, J. Ye, and Y. Liu, "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3656–3663.

[19] Y. Han, S. Wang, Y. Ren, C. Wang, P. Gao, and G. Chen, "Predicting station-level short-term passenger flow in a citywide metro network using spatiotemporal graph convolutional neural networks," *ISPRS International Journal of Geo-Information*, vol. 8, no. 6, p. 243, 2019.

[20] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 914–921.

[21] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, USA:, 1994, pp. 359–370.

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[23] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[24] K. Saleh, "Pedestrian trajectory prediction for real-time autonomous systems via context-augmented transformer networks," *Sensors*, vol. 22, no. 19, p. 7495, 2022.

[25] K. Saleh, A. Grigorev, and A.-S. Mihaita, "Traffic accident risk forecasting using contextual vision transformers," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 2086–2092.

[26] L. Bai, L. Yao, S. S. Kanhere, X. Wang, and Q. Z. Sheng, "Stg2seq: spatial-temporal graph to sequence model for multi-step passenger demand forecasting," in *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*. International Joint Conferences on Artificial Intelligence, 2019, pp. 1981–1987.

[27] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," *ICLR 2018*, 2018.

[28] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.