

SGRU: A High-Performance Structured Gated Recurrent Unit for Traffic Flow Prediction

1st Wenfeng Zhang

Shandong University

Jinan, China

202235331@mail.sdu.edu.cn

2nd Xin Li*

Shandong University

Jinan, China

lx@sdu.edu.cn

3rd Anqi Li

Shandong University

Jinan, China

2489316785@qq.com

4th Xiaoting Huang

Shandong University

Jinan, China

satinhuang@sdu.edu.cn

5th Ti Wang

China Unicom Smart City Research Institute

Beijing, China

wangti2@chinaunicom.cn

6th Honglei Gao

Shandong Cultural and Tourism Group Co., Ltd

Jinan, China

1348513@qq.com

Abstract—Traffic flow prediction is an essential task in constructing smart cities and is a typical Multivariate Time Series (MTS) Problem. Recent research has abandoned Gated Recurrent Units (GRU) and utilized dilated convolutions or temporal slicing for feature extraction, and they have the following drawbacks: (1) Dilated convolutions fail to capture the features of adjacent time steps, resulting in the loss of crucial transitional data. (2) The connections within the same temporal slice are strong, while the connections between different temporal slices are too loose. In light of these limitations, we emphasize the importance of analyzing a complete time series repeatedly and the crucial role of GRU in MTS. Therefore, we propose SGRU: Structured Gated Recurrent Units, which involve structured GRU layers and non-linear units, along with multiple layers of time embedding to enhance the model’s fitting performance. We evaluate our approach on four publicly available California traffic datasets: PeMS03, PeMS04, PeMS07, and PeMS08 for regression prediction. Experimental results demonstrate that our model outperforms baseline models with average improvements of 11.7%, 18.6%, 18.5%, and 12.0% respectively.

Index Terms—Traffic flow prediction, Multivariate time series, Gated recurrent units, Structured gated recurrent units

I. INTRODUCTION

The scenario of multivariate time series occurs in various domains of life. Researchers utilize historical weather data from different regions to predict future rainfall intensity [1]. Taxi data is used to help cities predict travel resources and reduce traffic congestion [2]. Exception monitoring is performed on various states in manufacturing systems and internet services [3]. Since the introduction of Graph Convolutional Networks (GCN) in 2017 [4], this method has been widely used in the field of Multivariate Time Series (MTS) for spatial semi-supervised and self-supervised learning. By interleaving one-dimensional convolution with gated linear units (GLU) and graph convolution, and appending an output layer after this “sandwich” structure [5], accurate prediction of traffic flow speed can be achieved. The combination of recurrent neural networks and GCN, with linear layers for

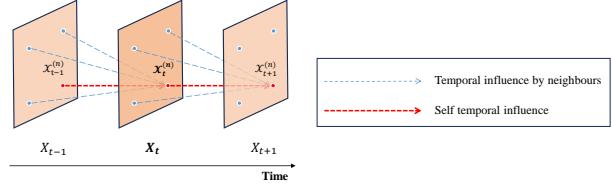


Fig. 1. Traffic Flow Prediction: a typical MTS problem

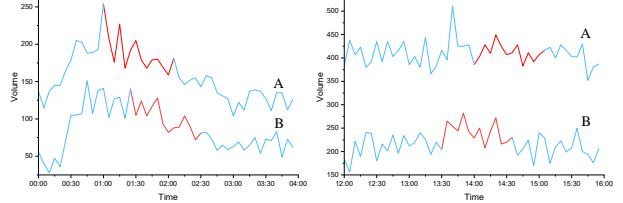


Fig. 2. Traffic waveforms of two adjacent nodes (they are similar with a lag of about 30 minutes)

output, has made significant contributions to accurate fitting of multivariate traffic flow data [6]. By incorporating attention mechanisms into GCN, it is possible to distinguish the importance of different nodes and utilize Gated Recurrent Units (GRU) for frequency domain feature extraction on long time series, leading to a substantial improvement in prediction accuracy [7]. The aforementioned work demonstrates the effectiveness of GCN in learning spatial features and its strong adaptability to non-Euclidean structured data.

A. Problem 1: Dilated convolutions break adjacent time steps

Focusing on the direction of traffic flow prediction (Figure 1,2), recent Temporal Convolutional Networks [8] and Graph WaveNet (GWNet) [9] have adopted the mechanism of GCN. However, in the frequency domain, they use dilated convolutions to extract sequence features, which leads to Problem 1: the neglect of the connection between adjacent time steps. The time series in the dataset are discretely

* Corresponding Author.

sampled, typically every few minutes, and the longer time interval makes the connection between adjacent time steps more important. The use of causal convolutions results in a large number of stacked convolutional layers, increasing the difficulty of training. Additionally, all the convolutional layers share the same convolutional kernel parameters, limiting the features already extracted by the previous layers [10]. Dilated convolutions are used to expand the global receptive field without linearly increasing the depth of convolutional layers and to expedite the convolutional operations. GWNet avoids the vanishing and exploding gradients caused by recurrent neural networks by employing dilated convolutions. The experiments [9] select a convolutional kernel size of 2 and an alternating dilation sequence of 1, 2, 1, 2, 1, 2, 1, 2, which achieves the best experimental results. This precisely demonstrates that when applying convolutional operations to the time domain, larger dilations (greater than or equal to 3) should not be used because abandoning crucial information from neighboring time steps leads to a decrease in model performance. Consequently, the experiment employs 8 layers of dilated convolutions to extract sequence features of length 13, significantly increasing the training parameters and time. The traditional dilated convolutions that ignore the connection between adjacent time steps are not a promising approach.

B. Problem 2: Time slices cause a weak of spatial connectivity

To enhance the connection between adjacent time steps, STGCN [11] utilizes GCN while sampling T time steps and treating every 3 time steps as a slice. The sliding window moves only 1 step at a time, generating sequences of length $T - 2$. GCN is used to extract spatial features from each time slice, followed by activation functions and linear transformations to restore the original dimensionality. This leads to Problem 2: although the fine-grained time slices largely aggregate the features of adjacent time steps, the lack of inter-slice correlation extraction through GCN results in a weak of spatial connectivity. All the slices are stacked together after GCN, and due to the high repetition in the time dimension, the linear layers require a large number of trainable parameters, extending the model's training time. In order to match each group of 3 time slices with their corresponding adjacency matrix, the trainable parameters of size $N \times N$ are expanded to $3N \times 3N$, nearly exponentially increasing the model size. However, the large number of training parameters is confined within a single slice containing only 3 time steps, greatly limiting the model's temporal receptive field. Therefore, dividing the sequence into time slices is not the optimal choice.

C. Our solution

To address the aforementioned issues, we propose a Structured Gated Recurrent Units (SGRU) network that treats long time series as a whole. By combining SGRU with Graph Convolutional Networks (GCN) and performing multiple feature

extractions on the same sequence, we effectively preserve the correlation information among adjacent time steps. Compared to the linear GRU combined with GCN [12], [13], SGRU utilizes non-linear layers to connect multiple GRUs, thereby enhancing the model's feature extraction capability. Additionally, prior to inputting the data into GRU, we apply multi-layer spatio-temporal embedding. The main contributions of our work are as follows:

- We propose a structured GRU layer that performs more comprehensive feature extraction on temporal data. Compared to a simple linearly connected GRU layer, SGRU has a smaller depth and faster fitting speed.
- We propose a multi-layer spatio-temporal embedding approach that maps high-dimensional features to low-dimensional features and embeds them before the recurrent network. This method enables effective analysis of heterogeneity among nodes.
- We conduct regression predictions on four publicly available California traffic datasets to evaluate the performance of our SGRU model compared to other models. The experimental results demonstrate that our approach generally outperforms the baselines. Additionally, we perform thorough ablation experiments on multiple datasets to verify the effectiveness of the key components of our model. Our code is publicly available on GitHub: <https://github.com/atxxfs/SGRU>

II. RELATED WORKS

Initially, the prediction of multivariate time series (MTS) was based on statistical methods such as Autoregressive Integrated Moving Average (ARIMA) [14]. Subsequently, with the introduction of Support Vector Machines (SVM) [15] by N. Zhang et al. in 2011 and Support Vector Regression (SVR) [16] by R. Chen et al. in 2015, traditional machine learning theories found their way into the realm of MTS. In recent years, the rapid advancement of deep learning has led to the emergence of numerous deep models focused on feature extraction techniques. In deep learning approaches, adjacency matrices are employed to represent the relationships between different nodes. Based on whether the adjacency matrix undergoes changes during model updates, it is categorized into static and dynamic adjacency matrices.

A. Static Adjacency Matrices

Prior to model training, a fixed and unchanging adjacency matrix is generated based on the distances between pairs of nodes. ASTGCN [17] integrates three types of attention spans - "hourly", "daily", and "weekly" - to dynamically capture spatiotemporal features using convolution. STGDN [18] constructs a multi-scale attention network, endowing the model with the capability to capture multi-level temporal dynamics. HGCN [19] employs dynamically interacting regional node adjacency matrices to focus on traffic features in the Central Business District (CBD) of a city. FOGS [20] adopts a first-order gradient-supervised method, converting dataset labels

from "flow" to "trend" before model training. The benefit of these approaches lies in the interpretability of adjacency matrix values, which align with the fixed geographical distances between observed nodes.

B. Dynamic Adjacency Matrices

Dynamic adjacency matrices are updated during backpropagation, either through random initialization or pre-definition. In DCRNN [21], Seq2Seq structures are employed to integrate spatiotemporal features, and recurrent networks are used to predict traffic data. STNN [22] introduces a novel module to simultaneously capture features in both temporal and spatial domains. ASTTN [23] separates different time steps of distinct nodes, scaling attention spans to a 1-hop spatial neighborhood to reduce complexity. DSTAGNN [24] calculates the proportion of traffic flow on a given day relative to other days using Wasserstein distance, generating spatiotemporal correlation graphs. The advantage of these methods lies in their adaptability to dynamic adjacency changes caused by congestion or unforeseen traffic incidents, aligning more closely with real-world scenarios. Therefore, we also utilize dynamic adjacency matrices.

III. PRELIMINARY

In the original sequence $\mathcal{X} \in \mathbb{R}^{T \times N \times D}$ where T represents the temporal duration, N represents the number of nodes, and D represents the feature dimension, if we denote the current time step as t and sample \mathcal{X} along the temporal dimension, we obtain a sliced tensor. For example, $\mathcal{X}_{t:t+F} \in \mathbb{R}^{(F+1) \times N \times D}$ where the temporal span is $F + 1$. Another example is $X_t \in \mathbb{R}^{N \times D}$, which represents the D features of N nodes at the current time step. We define the graph problem of traffic flow as $G = (V, \mathcal{E}, \mathcal{A}, \mathcal{X}_{t-P+1:t})$, where V represents N sensor nodes on the traffic network, and $|V| = N$. \mathcal{E} denotes the edges between the nodes. \mathcal{A} represents the adjacency matrix of G , with $\mathcal{A} \in \mathbb{R}^{N \times N}$. $\mathcal{X}_{t-P+1:t} \in \mathbb{R}^{P \times N \times D}$ describes the historical traffic flow observed by N sensors over the past P time steps. Our objective is to learn a prediction function $h(\cdot)$ that predicts the future traffic flow data $\mathcal{Y}_{t+1:t+F} \in \mathbb{R}^{F \times N \times D}$ for the next F time steps. This can be expressed as the following formula:

$$\mathcal{X}_{t-P+1:t} \xrightarrow{h(\cdot)} \mathcal{Y}_{t+1:t+F} \quad (1)$$

IV. METHODOLOGY

A. Preparation & Embedding

1) *Adaptive graph convolution*.: We adopt a method to construct the adaptive adjacency matrix using two spatial embedding matrices $E_1, E_2 \in \mathbb{R}^{N \times d}$. The vectors are randomly initialized without any prior knowledge and the parameters are updated during the learning process. The adjacency features of nodes are represented by multiplication:

$$A = \text{SoftMax}(\text{ReLU}(E_1 E_2^T)) \quad (2)$$

Two different spatial embedding matrices increase the length of the feature vectors for each node to $2d$, in order to

improve the model's fitting capability. By multiplying the two embedding matrices, an $N \times N$ matrix is obtained. The ReLU activation function filters out negative values, and the SoftMax function normalizes the matrix values, resulting in the adaptive adjacency matrix A .

2) *Multilayer spatiotemporal embedding*.: The time embedding vector and spatial embedding vector, $E_{space} \in \mathbb{R}^{N \times d'}$ and $E_{time} \in \mathbb{R}^{d' \times P}$, are randomly initialized. They are added to the original sequence using a broadcasting mechanism and dimension consistency is ensured through linear layers. $W_e \in \mathbb{R}^{d \times d'}$ and $b_e \in \mathbb{R}^{d'}$ represent the weight and bias of the linear layer, respectively. The introduction of d' is to flexibly change the dimension of the embedding matrix, resulting in $\tilde{\mathcal{X}} \in \mathbb{R}^{P \times N \times d'}$

$$\tilde{\mathcal{X}} = \mathcal{X} W_e + b_e + E_{space} + E_{time}^T \quad (3)$$

B. Temporal feature extraction.

To capture the temporal features of the complete time series and prevent gradient vanishing or exploding, we utilize a GRU recurrent network with GCN embedded inside. It is represented as:

$$\begin{aligned} LEFT_t &= A \left[\tilde{\mathcal{X}}_t W_s^{(1)} + b_s^{(1)}, h_{t-1} \right] \\ RIGHT_t &= I \left[\tilde{\mathcal{X}}_t W_s^{(2)} + b_s^{(2)}, h_{t-1} \right] \\ \tilde{A}_t &= [LEFT_t, RIGHT_t] W_a + b_a \end{aligned} \quad (4)$$

First, we obtain $\tilde{\mathcal{X}}_t W_s + b_s \in \mathbb{R}^{N \times H}$ through a linear transformation, where H represents the hidden layer dimension of the recurrent network. I is an $N \times N$ identity matrix, $W_a \in \mathbb{R}^{4H \times H}$, $b_a \in \mathbb{R}^{N \times H}$ represent the weights and biases of the linear transformation, respectively. $h_{t-1} \in \mathbb{R}^{N \times H}$ is the output of the previous time step's hidden layer. After obtaining the feature matrix with dimension transformation, denoted as $\tilde{A}_t \in \mathbb{R}^{N \times H}$, we then input it into the GRU:

$$\begin{aligned} z_t &= \sigma(\tilde{A}_t W_z + b_z) \\ r_t &= \sigma(\tilde{A}_t W_r + b_r) \\ C_t &= \tanh \left(\left[\tilde{X}_t, r_t \circ h_{t-1} \right] W_c + b_c \right) \\ h_t &= (1 - z_t) \circ C_t + z_t \circ h_{t-1} \end{aligned} \quad (5)$$

C. Structured GRUs network.

A single GRU has limited regression performance. To enhance the model's performance, we adopt a structured multi-GRUs network. Different GRUs extract temporal features of the same sequence, and the hidden layer outputs are aggregated, linearly transformed, and then outputted. The model structure is illustrated in Figure 3, and the comparison between Linear GRUs and SGRU is illustrated in Figure 4.

In SGRU, the output of GRU(a) and GRU(b), $h_t^{(a)}$ and $h_t^{(b)}$, respectively, pass through three linear units. $h_t^{(a)}$ additionally undergoes a sigmoid activation and a Hadamard product is applied to connect the two parts as follows:

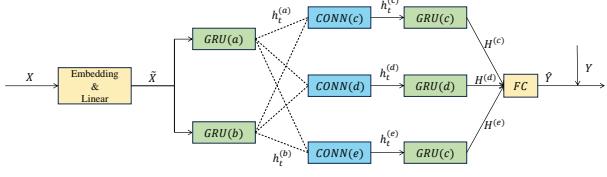


Fig. 3. Model structure of SGRU

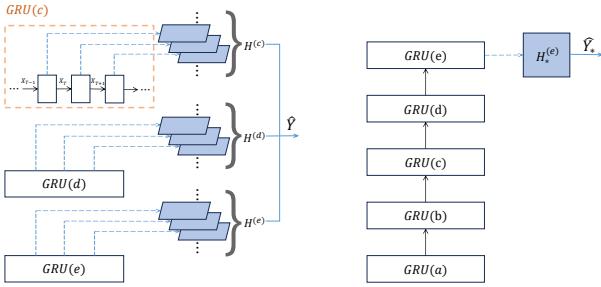


Fig. 4. Structural comparison between SGRU (left) and Linear GRUs (right)

$$\begin{aligned} h_t^{(c)} &= \sigma(h_t^{(a)}W_{ac} + b_{ac}) \circ (h_t^{(b)}W_{bc} + b_{bc}) \\ h_t^{(d)} &= \sigma(h_t^{(a)}W_{ad} + b_{ad}) \circ (h_t^{(b)}W_{bd} + b_{bd}) \\ h_t^{(e)} &= \sigma(h_t^{(a)}W_{ae} + b_{ae}) \circ (h_t^{(b)}W_{be} + b_{be}) \end{aligned} \quad (6)$$

The output results $h_t^{(c)}$, $h_t^{(d)}$, and $h_t^{(e)}$ initialize the hidden layers of GRU(c), GRU(d), and GRU(e), respectively.

The widely used linear GRU (shown in Figure 4, right) operates by feeding the temporal data into the first GRU, passing the output to the next GRU, and using the output of the last GRU for making predictions [13]. SGRU improves upon this structure (shown in Figure 3 and Figure 4, left) by introducing the Connection Module (CONN), which simultaneously preserves the nonlinear features from the output of GRU(a) and the linear features from the output of GRU(b), and merges them together. The final prediction includes the outputs of all three SGRUs, resulting in three times the output of the linear GRU and better performances.

D. Prediction and loss calculation.

The outputs of the last three GRUs at all time steps, which are $\mathcal{H}^{(c)}$, $\mathcal{H}^{(d)}$ and $\mathcal{H}^{(e)}$, are concatenated and passed through a fully connected (FC) layer to obtain the final result.

$$\begin{aligned} \mathcal{H} &= [h_{t-P+1}, h_{t-P+2}, \dots, h_t] \\ \hat{Y} &= FC([\mathcal{H}^{(c)}, \mathcal{H}^{(d)}, \mathcal{H}^{(e)}]) \end{aligned} \quad (7)$$

In order to predict the future states of N nodes over a time horizon of F time steps, The model loss is computed using the mean absolute error (MAE), and the model parameters are updated using back propagation with gradient descent.

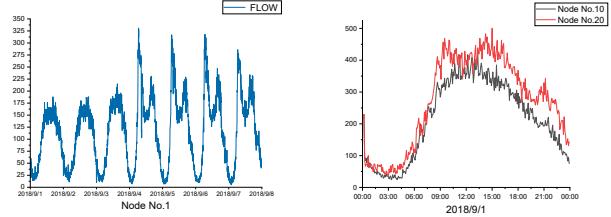


Fig. 5. PeMS03 visualization of nodes No.10 and No.20

The Adam optimizer is utilized to minimize the value of \mathcal{L} over multiple iterations.

$$\mathcal{L}(\hat{Y}, Y) = \frac{1}{F \times N} \sum_{f=1}^F \sum_{n=1}^N |\hat{Y}_f^{(n)} - Y_f^{(n)}| \quad (8)$$

V. EXPERIMENT

In this chapter, we evaluate the performance of SGRU on publicly available datasets and compare it with other models. Additionally, we conduct ablation experiments by removing certain important modules to verify their effectiveness by comparing the performance with the original model.

A. Datasets

We evaluate our model on four datasets from the Performance Measure System (PeMS) of the California highway network [25]:

- **PeMS03:** 358 nodes, 26,209 time steps, starting from September 1, 2018.
- **PeMS04:** 307 nodes, 16,992 time steps, starting from January 1, 2018.
- **PeMS07:** 883 nodes, 28,224 time steps, starting from May 1, 2017.
- **PeMS08:** 170 nodes, 17,856 time steps, starting from July 1, 2016.

The time series data for nodes 1, 10 and 20 of PeMS03 are visualized, revealing clear periodicity (left side of Figure 5). Additionally, there are noticeable similarities (before 8 a.m.) and differences (after 8 a.m.) among different nodes within the same day (right side of Figure 5).

Each dataset is preprocessed by linear interpolation to fill missing values, followed by standardization:

$$X' = \frac{X - \text{mean}(X)}{\text{std}(X)} \quad (9)$$

The data is sampled at 5-minute intervals, resulting in 288 time steps per day. The entire dataset was divided into training, validation, and testing sets using a ratio of 6:2:2.

B. Model Settings

To demonstrate the superiority of our model, we compare it with the following baseline models:

- **DCRNN** [21]: Uses a Seq2Seq structure to integrate spatial and temporal features and predict traffic data using recurrent networks.

TABLE I
PERFORMANCE COMPARISON TO OTHER BASELINES ON FOUR PUBLICLY AVAILABLE DATASETS.

Datasets		PeMS03			PeMS04			PeMS07			PeMS08		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Models	DCRNN	18.18	30.31	18.63	24.70	38.12	17.12	28.30	38.58	11.66	17.86	27.83	11.45
	STGCN	17.49	30.12	17.15	22.70	35.55	14.59	25.38	38.78	11.08	18.02	27.83	11.40
	GWNet	19.85	32.94	19.31	25.45	39.70	17.29	26.85	42.78	12.12	19.13	31.05	12.68
	ASTGCN	17.69	29.66	19.40	22.93	35.22	16.56	28.05	42.57	13.92	18.61	28.16	13.08
	STSGCN	17.48	29.21	16.78	21.19	33.65	13.90	24.26	39.03	10.21	17.13	26.80	10.96
	STFGNN	16.77	26.28	16.30	20.48	32.51	16.77	23.46	36.60	9.21	16.94	26.25	10.60
	STID	19.11	31.42	18.98	23.94	37.51	16.12	26.01	40.60	11.63	19.26	29.94	11.99
	STGNCDE	19.92	32.53	19.40	26.35	41.57	19.90	29.22	43.79	14.05	19.97	32.23	13.60
	STSSL	16.77	26.99	16.28	20.54	33.14	16.28	22.81	34.71	9.95	16.50	26.51	10.38
	GMAN	16.21	26.35	15.92	20.23	32.17	14.29	22.12	34.73	9.58	16.04	25.77	10.08
	StemGNN	17.31	28.00	17.53	21.06	33.87	15.91	23.10	35.67	10.25	17.00	26.31	11.55
	SGRU(Ours)	15.22	26.23	15.64	19.96	31.98	13.22	21.60	34.58	9.16	15.96	25.13	10.22
	Imp.	14.9%	10.9%	11.7%	12.0%	10.5%	18.6%	15.0%	11.1%	18.5%	10.6%	10.4%	12.0%

- **STGCN** [5]: Applies graph convolution in the spatial domain and two 1D convolutions in the temporal domain in a sandwich structure.
- **GWNet** [9]: Utilizes spatially adaptive matrix learning for adjacency transformation and applies dilated convolutions in the temporal domain.
- **ASTGCN** [17]: Incorporates multiple spatio-temporal attention mechanisms and captures spatio-temporal features using convolution operations.
- **STSGCN** [11]: Slices long sequences into smaller segments using a sliding time window and enhances local spatial perception on each segment.
- **STFGNN** [26]: Generates "time graphs" to complement the limitations of "spatial graphs" and learns hidden spatio-temporal dependencies.
- **STID** [6]: Applies spatio-temporal embedding to the original sequence, concatenates the embedding results, and outputs them using fully connected layers.
- **STGNCDE** [27]: The spatiotemporal data is processed using two neural controlled differential equations (NCDEs), and the results are then combined together.
- **STSSL** [10]: Learns spatio-temporal heterogeneity through self-supervised mechanisms and combines "heterogeneity-aware enhancement" to improve the model's robustness.
- **GMAN** [28]: Uses an encoder-decoder structure with attention mechanisms and incorporates GRU for spatio-temporal feature fusion.
- **StemGNN** [29]: Transforms the time domain to the frequency domain using discrete Fourier transform (DFT) and graph Fourier transform (GFT) to capture spatio-temporal dependencies simultaneously.

We set the number of historical time steps (P) to 12, the number of predicted time steps (F) to 12, the spatial embedding dimension (d) to 2, the GRU hidden dimension (H) to 64, and the batch size to 64. The experiments were conducted on an *Intel(R) Xeon(R) Gold 6226R* CPU and *NVIDIA GeForce RTX 3090* GPU, and multiple experiments were performed with random seeds ranging from 1 to 10, with the average

results being reported.

C. Performance Evaluation

We selected three evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Lower values indicate better model performance. The experimental results are presented in Table 1, where we calculated the average values of 11 baseline models across the three evaluation metrics. SGRU consistently achieved improvements of over 10%, with a particularly significant increase of 18.6% in MAPE on the PeMS04 dataset. Apart from the MAPE metric on the PeMS08 dataset, where SGRU performed slightly worse than GMAN, SGRU outperformed all other models in terms of all evaluation metrics.

GMAN demonstrated the second-best performance among the models. GMAN incorporates a mechanism of spatial-temporal attention with GRU, and its output results are combined with L ST-Attention blocks. However, each block considers the correlation between t_j and t time steps, resulting in excessive granularity in the temporal dimension and leading to time step repetition. As a result, the experimental results of GMAN were not as good as those of SGRU.

D. Ablation Study

To validate the effectiveness of the important modules in SGRU, we introduce the following variants:

- **simple**: The original model with 5 linear GRUs and no multi-layer spatiotemporal embedding.
- **w/ st-emb**: Adding multi-layer spatiotemporal embedding to the simple model.
- **w/ struct**: Replacing the 5 linear GRUs in the simple model with structured GRUs.

Due to the large size of the PeMS07 dataset, we conducted experiments sequentially on the PeMS03, PeMS04, and PeMS08 datasets. The evaluation metric was MAE. We varied the value of the future prediction time steps, denoted as F, which corresponded to F=3,6,9,12 (corresponding to 15 minutes, 30 minutes, 45 minutes, and 60 minutes, respectively). The results are shown in Figure 6 and Table 2.

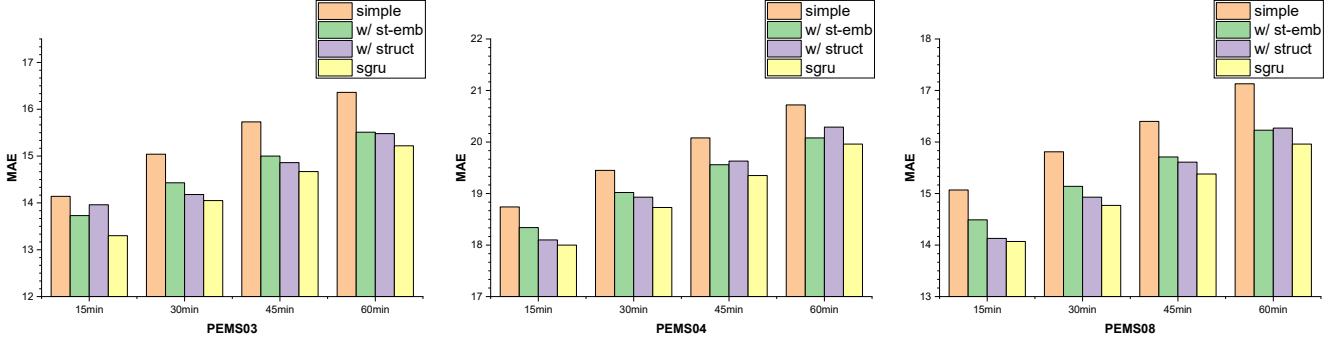


Fig. 6. MAE of SGRU and its 3 variants

TABLE II
MAE OF SGRU AND ITS 3 VARIANTS

		15min	30min	45min	60min
PeMS03	simple	14.14	15.04	15.73	16.36
	w/ st-emb	13.73	14.43	15.00	15.51
	w/ struct	13.96	14.18	14.86	15.48
	sgru	13.30	14.05	14.67	15.22
PeMS04	simple	18.74	19.45	20.08	20.72
	w/ st-emb	18.34	19.02	19.56	20.08
	w/ struct	18.10	18.93	19.63	20.29
	sgru	18.00	18.73	19.35	19.96
PeMS08	simple	15.07	15.81	16.40	17.13
	w/ st-emb	14.49	15.14	15.71	16.23
	w/ struct	14.13	14.93	15.61	16.27
	sgru	14.07	14.77	15.38	15.96

The simple model performs the worst, the addition of multi-layer spatiotemporal embedding improves its performance, and the use of structured GRUs significantly enhances the performance. SGRU achieves the best performance overall, confirming the contribution of both important modules to the final model.

VI. CONCLUSION

In order to achieve a better fitting performance on traffic datasets and avoid the loss of adjacent data caused by dilated convolutions or the loose connections between temporal slices, as well as the excessive bulkiness within each slice, we propose a Structured Gated Recurrent Units (SGRU) network. After completing the spatio-temporal embedding, SGRU employs two parallel GRUs to extract features from the same time series. These features are then transformed by non-linear layers and fed into three GRUs. Finally, the hidden layer outputs are concatenated, passed through fully connected layers, and used for regression prediction. We conducted comparative experiments on four publicly available datasets: PeMS03, PeMS04, PeMS07, and PeMS08. Compared to the average performance of baseline models, our approach achieved improvements of 11.7%, 18.6%, 18.5%, and 12.0% respectively. Furthermore, ablation experiments demonstrated a clear advantage of SGRU over linear GRU.

ACKNOWLEDGMENT

This research is sponsored by the National Key R&D Program of China (2021YFB1407004, 2019YFA0709401) and the Key Program of Shandong Province (2020CXGC010904, 2021SFGC04010401, 2023CXGC010801).

REFERENCES

- [1] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.
- [2] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, "Deep multi-view spatial-temporal network for taxi demand prediction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.
- [3] Z. Li, Y. Zhao, J. Han, Y. Su, R. Jiao, X. Wen, and D. Pei, "Multivariate time series anomaly detection and interpretation using hierarchical intermetric and temporal embedding," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 3220–3230.
- [4] H. Zhang, G. Lu, M. Zhan, and B. Zhang, "Semi-supervised classification of graph convolutional networks with laplacian rank constraints," *Neural Processing Letters*, pp. 1–12, 2021.
- [5] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," *arXiv preprint arXiv:1709.04875*, 2017.
- [6] Z. Shao, Z. Zhang, F. Wang, W. Wei, and Y. Xu, "Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 4454–4458.
- [7] J. Bai, J. Zhu, Y. Song, L. Zhao, Z. Hou, R. Du, and H. Li, "A3t-gcn: Attention temporal graph convolutional network for traffic forecasting," *ISPRS International Journal of Geo-Information*, vol. 10, no. 7, p. 485, 2021.
- [8] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
- [9] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," *arXiv preprint arXiv:1906.00121*, 2019.
- [10] J. Ji, J. Wang, C. Huang, J. Wu, B. Xu, Z. Wu, J. Zhang, and Y. Zheng, "Spatio-temporal self-supervised learning for traffic flow prediction," *arXiv preprint arXiv:2212.04475*, 2022.
- [11] 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, 2020, pp. 914–921.
- [12] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," *Advances in neural information processing systems*, vol. 33, pp. 17804–17815, 2020.

- [13] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-gcn: A temporal graph convolutional network for traffic prediction," *IEEE transactions on intelligent transportation systems*, vol. 21, no. 9, pp. 3848–3858, 2019.
- [14] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results," *Journal of transportation engineering*, vol. 129, no. 6, pp. 664–672, 2003.
- [15] N. Zhang, Y. Zhang, and H. Lu, "Seasonal autoregressive integrated moving average and support vector machine models: prediction of short-term traffic flow on freeways," *Transportation Research Record*, vol. 2215, no. 1, pp. 85–92, 2011.
- [16] R. Chen, C.-Y. Liang, W.-C. Hong, and D.-X. Gu, "Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm," *Applied Soft Computing*, vol. 26, pp. 435–443, 2015.
- [17] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 922–929.
- [18] X. Zhang, C. Huang, Y. Xu, L. Xia, P. Dai, L. Bo, J. Zhang, and Y. Zheng, "Traffic flow forecasting with spatial-temporal graph diffusion network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 17, 2021, pp. 15 008–15 015.
- [19] K. Guo, Y. Hu, Y. Sun, S. Qian, J. Gao, and B. Yin, "Hierarchical graph convolution network for traffic forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 1, 2021, pp. 151–159.
- [20] X. Rao, H. Wang, L. Zhang, J. Li, S. Shang, and P. Han, "Fogs: First-order gradient supervision with learning-based graph for traffic flow forecasting," in *Proceedings of International Joint Conference on Artificial Intelligence, IJCAI*. ijcai.org, 2022.
- [21] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," *arXiv preprint arXiv:1707.01926*, 2017.
- [22] S. Yang, J. Liu, and K. Zhao, "Space meets time: Local spacetime neural network for traffic flow forecasting," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 817–826.
- [23] A. Feng and L. Tassiulas, "Adaptive graph spatial-temporal transformer network for traffic forecasting," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3933–3937.
- [24] S. Lan, Y. Ma, W. Huang, W. Wang, H. Yang, and P. Li, "Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting," in *International conference on machine learning*. PMLR, 2022, pp. 11 906–11 917.
- [25] C. Chen, K. Petty, A. Skabardonis, P. Varaiya, and Z. Jia, "Freeway performance measurement system: mining loop detector data," *Transportation Research Record*, vol. 1748, no. 1, pp. 96–102, 2001.
- [26] M. Li and Z. Zhu, "Spatial-temporal fusion graph neural networks for traffic flow forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 4189–4196.
- [27] J. Choi, H. Choi, J. Hwang, and N. Park, "Graph neural controlled differential equations for traffic forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 6367–6374.
- [28] C. Zheng, X. Fan, C. Wang, and J. Qi, "Gman: A graph multi-attention network for traffic prediction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 1234–1241.
- [29] D. Cao, Y. Wang, J. Duan, C. Zhang, X. Zhu, C. Huang, Y. Tong, B. Xu, J. Bai, J. Tong *et al.*, "Spectral temporal graph neural network for multivariate time-series forecasting," *Advances in neural information processing systems*, vol. 33, pp. 17 766–17 778, 2020.