Spatial-temporal Transformer Network with Self-supervised Learning for Traffic Flow Prediction

Author Name
Affiliation
pcchair@ijcai-22.org

Abstract

Traffic flow prediction plays a critical role in improving the quality, security, and efficiency of Intelligent Transportation Systems (ITS). As a spatial-temporal problem, accurate prediction requires modeling spatial and temporal characteristics. Although the spatial dependencies have been considered in existing works, they focus more on the local perspective, performing poorly in global dependencies. In addition, a simple property: temporal similarity in traffic flow data is overlooked by existing work. They focus on the high-level temporal semantic (e.g., periodicity), and fail to consider the similarities between future data and historical observations as a strong prior knowledge for the prediction. At last, the data-hungry paradigm is an inevitable curse for current deep learning methods, while existing data is limited and collecting more data is expensive. How to effectively explore spatial-temporal features with limited data has become an inescapable issue. To alleviate these limitations, we propose Spatial-temporal Transformer Network with Self-supervised Learning (ST-TSNet). ST-TSNet uses a Pre-Conv Block and vision transformer to learn the spatial dependencies in both local and global contexts. Furthermore, a skip connection from the input of historical records to the output prediction is introduced to utilize similar patterns to facilitate prediction. Finally, a self-supervised strategy called stochastic argumentation is proposed to explore spatial-temporal representations to benefit the prediction task. Extensive experiments are conducted on two open datasets: TaxiBJ and TaxiNYC, demonstrating the effectiveness of ST-TSNet. The codes will be available upon acceptance.

1 Introduction

With the rapid growth of traffic sensors deployed, a massive amount of traffic flow data is collected, motivating researchers to apply deep learning methods to model the traffic flow data for traffic management. Traffic flow data is a 2D spatial tensor that contains the traffic flow of an entire city. An accurate prediction of future traffic flow data depends on modeling the relevant information from the previous observations and nearby neighbors. This problem can be considered from the spatial and temporal perspectives. From the spatial perspective, learning the local spatial correlations is essential as traffic volume is most influenced by its nearest neighbors. On the other hand, the global dependencies are equally significant. In real-world scenarios, two distant regions may be strongly correlated in their traffic distributions as they feature the same functionality (e.g., transportation hub). However, existing works [Zhang *et al.*, 2017; Guo *et al.*, 2019; Fang *et al.*, 2019] adopt the convolutional layers as their backbone, which yields a small receptive field, introducing short-range bias. Thus they perform well in the local context while hindering in global dependencies. Recently Vision transformer (ViT) [Dosovitskiy *et al.*, 2021] has shown impressive performance in computer vision, due to its innate power at extracting non-local features. We are motivated to apply ViT to make up the need for learning long-range spatial dependencies.

From the temporal perspective, many works have been proposed to extract complex temporal patterns, e.g., daily and weekly periodicity [Zhang *et al.*, 2017; Guo *et al.*, 2019]. However, we argue that a simple temporal characteristic: temporal similarity is overlooked. Traffic flow data are innately continuous time series generally without abrupt changes, showing many similarities in adjacent frames. As depicted in the time series of Fig.1, the ratio of current traffic flow to the previous one (shown in blue line) floats up and down from a fixed ratio of 1 as the traffic flow (shown in orange line) periodically evolves. This means that adjacent traffic flow snapshots have a close value and exhibit similar distribution. Similarly, for the traffic flow prediction task, the future data is similar to the historical records. An intuitive idea is to use historical observations as the base prediction for future data. Such an idea provides a strong and reasonable prior knowledge that forces the model to predict the future data directly based on the historical records themselves instead of the temporal patterns extracted from them. However, existing methods [Zhang *et al.*, 2016a; Guo *et al.*, 2019] that process the historical data for high-order temporal characteristics (e.g., periodicity) distort its semantic information, resulting in the loss of such simple property.

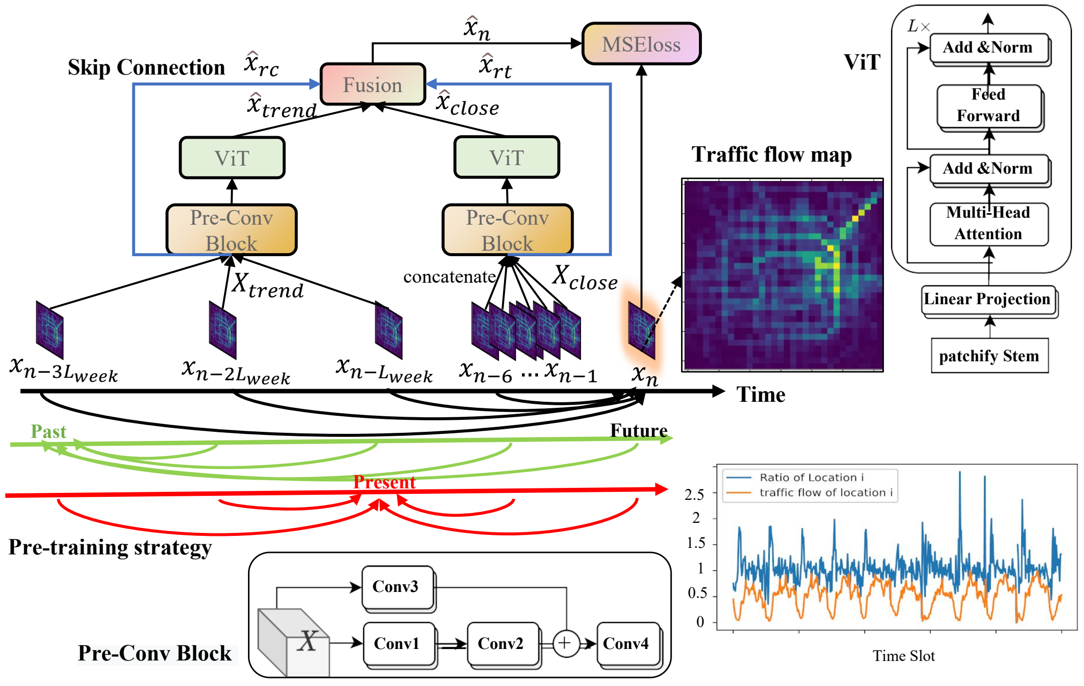

Figure 1: The overall architecture of Spatial-temporal Transformer Network with Self-supervised Learning (ST-TSNet). The three time axes illustrate our pre-training strategy. The time series at the bottom shows the periodicity of traffic flow data; the blue line denotes the ratio, and the orange line denotes the normalized traffic flow observations. The figure reveals that as the traffic flow periodically changes, the ratio floats up and down from a fixed value 1.

Finally, existing works suffer from a common problem: The data-hungry paradigm. It has become common sense that deep learning models require a sufficiently large amount of data to perform well, which is also an inescapable curse in the field of traffic flow prediction. However, existing available benchmarks are very small, typically several times less than the data from the mainstream field (e.g., computer vision). For example, as a recognized benchmark in traffic flow prediction task, TaxiBJ has 20,016 samples, which is only 30% of CIFAR-10 (an image recognition dataset that contains 60,000 samples). In addition, in the real world, collecting sufficient data is time-consuming (more than a year) and extremely expensive as it involves the deployment and maintenance of the sensors in an entire city. How to effectively explore spatial-temporal features under limited data has become a common problem faced by every work in this field. The solution of existing works is proposing models based on expert knowledge (e.g., short-range bias). However, it requires human expertise and violates the nature of deep learning: learning from data. There is countless domain knowledge buried in the data that humans can not list exhaustively. A more general solution is to automatically explore the spatial-temporal representations in the data by generating supervision signals with the input data itself.

Driven by these analyses and solutions, we propose a novel framework called Spatial-temporal Transformer Network with Self-supervised Learning (ST-TSNet). ST-TSNet consists of a Pre-Conv Block and ViT for learning spatial correlations in both local and global contexts. In addition, we directly connect the historical data to the output to utilize the historical data as the base predictions and employ well-designed modules to adjust it based on learned complex patterns. Lastly, a self-supervised task named stochastic augmentation is proposed to pre-train our ST-TSNet to learn spatial-temporal representations and fine-tune them to benefit the prediction task.

Our contributions are summarized as follows.

- We propose a novel framework Spatial-temporal Transformer Network with Self-supervised Learning (ST-TSNet) to capture spatial-temporal features.
- We employ a simple yet effective skip connection strategy, plugged into ST-TSNet, to utilize the temporal similarities in traffic flow data.
- We introduce self-supervised learning to our framework and design a pre-training task called stochastic augmentation to alleviate the limited data issue in traffic flow prediction task.
- We conduct extensive experiments on two benchmarks (TaxiBJ and TaxiNYC) to evaluate the effectiveness of our methods and the results show that our ST-TSNet outperforms state-of-the-art methods.

2 Related Work

Traffic Flow Prediction. Most mainstream traffic prediction methods fall into one of two classes: statistical methods or deep learning methods. Statistical methods include auto-regressive integrated moving average (ARIMA) [Williams and Hoel, 2003], Kalman filtering [hua Guo *et al.*, 2014] and historical average. They often require strong and trivial theoretical assumptions, which violates the nonlinearity of traffic flow data, thus having poor performance in the real world. Recent advances have witnessed the impressive capacity of deep learning to extract nonlinear features from big data [Salakhutdinov, 2014]. Many researchers are inspired to apply deep learning to traffic flow prediction task. Existing deep learning methods are based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [Lv *et al.*, 2015]. ST-ResNet [Zhang *et al.*, 2017] first employs the CNNs with residual connections to learn the spatial dependencies and construct historical data into different branches according to the temporal semantics to learn temporal features. Similar ideas are adopted by subsequent works [Guo *et al.*, 2019; Chen *et al.*, 2018], where they use 3D convolution to learn the spatial-temporal dependencies. Moreover, RNN-based models [Zhao *et al.*, 2017; Fiorini *et al.*, 2020] are inspired to use convolutional layer to capture spatial features and sequential hierarchy (e.g., LSTM and GRU) to extract temporal patterns. However, they are time-consuming as they make predictions step by step and may suffer gradient vanishing or explosion when capturing long-range sequences [Xu *et al.*, 2018]. To alleviate the problem, [Xu *et al.*, 2018; Fiorini *et al.*, 2021] discard the recurrent chain structure and employ Multiplicative Cascade Unit (CMU) with autoencoders while preserving the convolutional layers for learning spatial features. Methods used by existing works can be considered from spatial and sequential perspectives. From the spatial perspective, convolutional layers are the mainstream, including 2D and 3D convolution. From the sequential perspective, there are many choices, including RNN, GRU, LSTM and CMU. Most existing works are a combination of these methods. In summary, existing methods that based on CNNs suffer from short-range bias as the small receptive field limits their capacity to extract global dependencies.

Self-supervised Learning. Self-supervised learning is a great way to extract training signals from massive amounts of unlabelled data and to learn general representation to facilitate downstream tasks where the data is limited. To generate supervision information from data, a general strategy is to define pre-training tasks for models to solve [Zhang *et al.*, 2016b; Pathak *et al.*, 2016], during which models learn semantic representations and transfer them to downstream tasks to improve performance and robustness. Many works in computer vision have defined various tasks based on heuristics [Misra and van der Maaten, 2020; Noroozi and Favaro, 2016]. For example, [Gidaris *et al.*, 2018] learns visual representations by predicting the image rotations. In natural language processing (NLP), masked language modeling, e.g., Bert [Devlin *et al.*, 2019] and GPT-3 [Brown *et al.*, 2020] have shown to be excellent for pre-training language models. They mask a portion of the input sequence and train models to predict the missing content with the rest. Such methods are effective for learning semantic correlations of elements within a sequence, e.g., sentence. The traffic flow data can also be viewed as a sequence temporally, while the effectiveness of self-supervised learning remains unexplored in traffic flow prediction task.

3 Methods

3.1 Problem Formulation

We partition the traffic flow of a city into an image-like grid map, as shown in the traffic flow map of Fig.1, where each grid denotes a region. The value of a grid denotes the traffic flow (inflow or outflow). The device deployed at a region will periodically record the number of people arriving at and departing from the location to collect the inflow and outflow. The traffic flow map of the entire city at time t is noted as $x_t \in \mathbb{R}^{2\times H\times W}$, where 2 refers to the inflow and outflow, and H and W denote the map height and width, respectively. The purpose of traffic flow prediction is to predict x_n given historical traffic flow records $X_{his} = \{\mathbf{x}_t \mid t = 0, \ldots, n-1\}$. As shown in Fig.1, the historical data is summarized into two categories in the time axis: Closeness sequence $X_{close} = \{X_{n-1}, X_{n-2}, \cdots, X_{n-(d_c-1)}, X_{n-d_c}\} \in \mathbb{R}^{2\times d_c\times H\times W}$ is a concatenation of recent historical data where d_c is the length of closeness sequence. Trend sequence $X_{trend} = \{X_{n-L_{week}}, X_{n-2\cdot L_{week}}, \cdots, X_{n-d_t\cdot L_{week}}\} \in \mathbb{R}^{2\times d_t\times H\times W}$ is a concatenation of periodic historical data from the past few weeks, where d_t is the length of trend sequence, L_{week} is the number of intervals within a week.

3.2 Spatial-temporal Transformer Network

Overall, we employ a symmetric structure for X_{trened} and X_{close}: A Pre-Conv Block followed by a ViT with two shortcuts (i.e., two blue lines shown in Fig.1) from the input to the fusion layer. In the end, fusion layer adaptively merges four components (two residual components $\hat{X}_{rc}$ and $\hat{X}_{rt}$, two outputs $\hat{X}_{close}$ and $\hat{X}_{trend}$) to generate prediction $\hat{x}_n$.

Pre-Conv Block. The traffic flow in a region is highly relevant to its nearby locations. We design a Pre-Conv Block for such short-range dependencies. As illustrated in Fig.1, Conv1 and Conv2 are the main convolutional layers to capture short-range dependencies. Thus, we employ a small kernel size (i.e., 3×3) which gives the receptive field of 5. Such design ensures the Pre-Conv Block only captures the local dependencies at most in 5×5 regions. The short-range dependencies are well-captured by the Pre-Conv Block while leaving the long-range features to the vision transformer. Inserting CNNs before ViT has shown to be effective in strengthening the capacity of ViT [Hassani *et al.*, 2021]. Conv3 is the residual shortcut, employing 64 kernels with size 1×1, which adds up to the main branch as a residual component. For Conv1, Conv2, and Conv3, we use a large number of kernels (e.g., 64). Conv4 uses a much smaller number of kernels with a size of 1×1. By enlarging and then reducing the number of channels, Pre-Conv Block can learn various spatial-temporal dependencies and then refine them into a compact feature map.

Vision transformer. We apply vision transformer (ViT) [Dosovitskiy *et al.*, 2021] after the Pre-Conv Block to capture the global dependencies, as shown in the right of Fig.1. ViT is comprised of two main components: "Patchify" stem and transformer encoder. "Patchify" stem spatially splits the input feature map into non-overlap $p \times p$ patches and linearly projects patches into tokens. Each token contains the information of a patch of regions. Then the tokens are fused with learnable positional encoding to preserve the 2D positional information and fed into transformer encoder. The encoder utilizes a multi-head self-attention mechanism to model the long-range dependencies followed by a layer normalization and residual connection (Add & Norm) to the next sub-layer, where a Feed Forward Network (FFN) and another Add & Norm are employed to further process the tokens. Lastly, the tokens are averaged and then linearly transformed to generating output:$\hat{X}_{close}$ and $\hat{X}_{trend}$.

Skip Connection. Skip Connection are employed to transfer similar patterns from the historical observations to the output as the base prediction. To preserve the original similar patterns in historical data, we directly connect input X_c and X_t to the fusion layer, as shown in the blue line of Fig.1. Before connecting, we aggregate historical input data in the time dimension to match the shape. For two historical sequences $X_{close} \in \mathbb{R}^{2\times d_c \times H \times W}$ and $X_{trend} \in \mathbb{R}^{2\times d_t \times H \times W}$, we compute:

$$\hat{X}_{rc} = f(X_c) \in \mathbb{R}^{2\times 1 \times H \times W}, \tag{1}$$

$$\hat{X}_{rt} = f(X_t) \in \mathbb{R}^{2\times 1 \times H \times W}, \tag{2}$$

where $\hat{X}_{rc}$ and $\hat{X}_{rt}$ are the two residual components. $f(\cdot)$ is an aggregation function $\mathbb{R}^{2\times D \times H \times W} \rightarrow \mathbb{R}^{2\times 1 \times H \times W}$, where D denotes the length of historical data sequence. Here we use a summation function. Finally, the two residual components will be fused in the fusion layer, introduced next.

Fusion Layer. The degree of influence of the four components (i.e., two outputs $\hat{X}_{close}$, $\hat{X}_{trend}$ and two residual components $\hat{X}_{rc}$, $\hat{X}_{rc}$) is different, and the influence in different regions also varies. Therefore, to dynamically calibrate their contributions, we follow [Zhang *et al.*, 2018] to use a parametric-matrix-based fusion method, where the parameter matrices are learned from historical data. Formally,

$$\begin{aligned}\hat{X}_{pred} =& w_c \cdot \hat{X}_{close} + w_t \cdot \hat{X}_{trend} + \\ & w_{rc} \cdot \hat{X}_{rc} + w_{rt} \cdot \hat{X}_{rt},\end{aligned} \tag{3}$$

where $\cdot$ denotes element-wise multiplication, w is the learnable parameter that measures the influence of each component.

3.3 Self-supervised Learning with Stochastic Augmentation

A training sample, including input historical data (X_{close} and X_{trend}) and the target x_n, is sampled by temporal characteristics (e.g., closeness and periodicity), showing semantic correlations between each snapshot. As shown in the first axe (black) in Fig.1, the supervised training utilizes the historical records to predict the future data (past&present-to-future) to learn the correlations from history to future. We argue that such directional training limits the quality of learned spatial-temporal patterns. Because each snapshot in a training sample shows correlations with the rest according to their temporal semantic, the supervised training only learns one type (i.e., past&present-to-future). Our pre-training task stochastic argumentation aims to implement two other types of prediction: future&present-to-past (green axe) and past&future-to-present (red axe) to explore general spatial-temporal representations to facilitate the supervised training. Even though

Algorithm 1 The pre-training procedure with stochastic augmentation.

```
Input: MASA model: $f_\theta$, closeness data: $X_{close}$, trend data:
    $X_{trend}$, and predicted future data: $x_n$.
Output: pre-trained MASA model.
repeat
    $X_{group} \leftarrow X_{close} \cap X_{trend} \cap x_n$
    target $\alpha \leftarrow RandomSampling(X_{group})$
    Remaining snapshots $\Omega \leftarrow X_{group} - \alpha$

    pre-trained data $(\Omega, \alpha)$
    predictions $\hat{y} \leftarrow f_\theta(\Omega)$
    loss $\leftarrow MSELoss(\hat{y}, \alpha)$
    $backprop(loss)$
    update $f_\theta$
until stop criteria is met;
```

predicting the past based on the future is meaningless in the real world, the semantic correlations are shared by these predictions and can be transferred to benefit the supervised training. To implement multi-directional predictions, stochastic argumentation randomly predicts a snapshot from the rest of the sample. Such a context-based pre-training task for learning the sequential semantic is similar to Bert [Devlin *et al.*, 2019]. The pre-training strategy is depicted in Algorithm 1. We define a group as the union of closeness data, trend data, and predicted ground truth: $X_{group} = X_{close} \cap X_{trend} \cap x_n$. Then we randomly sample one snapshot as the target α and the rest data $\Omega = X_{group} - \alpha$ as the input, constructing pre-training data (Ω, α) to pre-train our model. It is worth noting that although our task lost the information of temporal order due to the random predictions, the spatial dependencies are not affected and thus reinforced by the multi-directional predictions.

Loss Function. We use Mean Square Error (MSE) as the objection function and gradient descent to optimize it during the training and pre-training.

4 Experiments

4.1 Dataset and Evaluation

Dataset. Our experiments are based on two traffic flow datasets: TaxiBJ and TaxiNYC. Additional external data including DayOfWeek, Weekday/Weekend, holidays, and meteorological data (i.e., temperature, wind speed, and weather) are processed into a one-hot vector. There are 20,016 constructed samples in TaxiBJ and 41,856 in TaxiNYC.

- TaxiBJ [Zhang *et al.*, 2018]: TaxiBJ is the citywide crowd flow dataset collected every half hour in Beijing. Based on the geographic area of Beijing, TaxiBJ is split into 32×32 regions.
- TaxiNYC [Fiorini *et al.*, 2021]: TaxiNYC is the taxi trip record dataset collected every one hour in New York City. TaxiNYC is divided into 16×8 regions based on the longitude and latitude of New York City[1].

[1]The raw records are available at the NYC government website. A processed version for experiments is available at github

Model	**TaxiBJ**			**TaxiNYC**		
	RMSE	MAPE (%)	APE	RMSE	MAPE (%)	APE
HA	40.93	30.96	6.77E+07	164.31	27.19	7.94E+05
ST-ResNet [Zhang *et al.*, 2018]	17.56±0.91	15.74±0.94	4.81E+07±3.03E+05	35.87±0.60	22.52±3.43	6.57E+05±1.00E+05
MST3D [Chen *et al.*, 2018]	21.34±0.55	22.02±1.40	4.81E+07±3.03E+05	48.91±1.98	23.98±1.30	6,98E+05±1.34E+04
ST-3DNet [Guo *et al.*, 2019]	17.29±0.42	15.64±0.52	3.43E+07±1.13E+06	41.62±3.44	25.75±6.11	7.52E+05±1.78E+05
3D-CLoST [Fiorini *et al.*, 2020]	17.10±0.23	16.22±0.20	3.55E+07±4.39E+05	48.17±3.16	22.18±1.05	6.48E+05±3.08E+04
STAR [Wang and Su, 2019]	16.25±0.40	15.40±0.62	3.38E+07±1.36E+06	36.44±0.88	25.36±5.24	7.41E+05±1.53E+05
PredCNN [Xu *et al.*, 2018]	17.42±0.12	15.69±0.17	3.43E+07±3.76E+05	40.91±0.51	25.65±2.16	7.49E+05±6.32E+04
STREED-Net [Fiorini *et al.*, 2021]	**15.61±0.11**	14.73±0.21	3.22E+07±4.51E+05	36.22±0.72	20.29±1.48	5.93E+05±4.31E+04
ST-TSNet (ours)	16.04±0.08	**14.63±0.05**	**3.20E+07±1.05E+5**	**34.34±0.32**	**15.68±0.09**	**4.58E+05±2.52E+03**

Table 1: Performance comparison of different methods on TaxiBJ and TaxiNYC.

Variant	TaxiBJ			TaxiNYC		
	RMSE	MAPE (%)	APE	RMSE	MAPE (%)	APE
ViT	20.16	34.68	7.60E+07	51.82	96.52	2.12E+08
ViT + SC	17.12±0.35	15.56±0.29	3.41E+07±6.29E+05	57.45±5.39	22.99±2.59	6.71E+07±7.57E+05
PC + SC	19.17±0.05	29.16±1.14	6.39E+07±2.50E+06	37.36±0.32	49.24±1.94	1.08E+08±4.25E+06
ViT + PC	16.34±0.21	14.70±0.13	3.22E+07±2.86E+05	37.29±2.88	16.83±0.24	4.91E+07±7.11E+04
ViT + PC + SC	16.14±0.16	14.62±0.06	3.20E+07±1.38E+05	34.87±0.39	16.18±0.20	4.72E+07±5.71E+04
ViT + PC + SC + SA	16.07±0.06	14.68±0.08	3.22E+07±1.72E05	34.47±0.23	15.90±0.08	4.64E+07±2.43E+04
ST-TSNet (w Ext)	**16.04±0.08**	**14.63±0.05**	**3.21E+07±1.05E+05**	**34.34±0.32**	**15.68±0.09**	**4.58E+07±2.52E+05**

Table 2: Ablation study of sub-modules in ST-TSNet.

Evaluation Metric. We adopt three commonly used metric: Rooted Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Absolute Percentage Error (APE). We follow previous works that compute the metrics on traffic flow value that is larger than 10 to ensure a fair comparison. We conducted experiments ten times for reliable results and present the results in the mean±std format.

4.2 Implementation Details

Min-Max normalization is applied in our experiments to scale the data to range $[-1, 1]$ and denormalize the predicted target back to the original value. We split the last 28 days as the test set for both datasets, and the remaining are regarded as training data. During training, we select 90% of the training data for training models and the remaining 10% is the validation set to early-stop our training algorithm. Our model is implemented and trained by PyTorch[2]. We use Adam [Kingma and Ba, 2015] as the optimizer with a learning rate of 0.001 for TaxiBJ and 0.005 for TaxiNYC. Cosine learning rate decay is employed to adjust the learning rate at each iteration. The batch size is 128 for both TaxiBJ and TaxiNYC. We run our model for 600 epochs on TaxiBJ and 800 epochs on TaxiNYC. Our ViT has two blocks, and the patch size is set to $(8, 8)$; the token dimension is set to 128; the number of attention heads is 2; the size of FFN is 512.

4.3 Quantitative Comparison

Table 1 shows the comparing results against the state-of-the-art methods. We compare our ST-TSNet with the following baselines: HA, ST-ResNet [Zhang *et al.*, 2018], MST3D [Chen *et al.*, 2018], ST-3DNet [Guo *et al.*, 2019], 3D-CLoST [Fiorini *et al.*, 2020], STAR [Wang and Su, 2019], and PredCNN [Xu *et al.*, 2018]. The baseline results are adopted from [Fiorini *et al.*, 2021].

[2]https://pytorch.org/

On TaxiBJ, our method exceeds the SOTA STREED-Net in terms of MAPE and APE and achieves comparable results in RMSE. While on TaxiNYC, our method significantly outperforms the SOTA ST-ResNet across all metrics by a fair margin (1.53 RMSE, 4.61 MAPE, and 1.35E+05 APE improvement).

ST-TSNet has a more significant performance improvement on TaxiNYC than TaxiBJ. Because TaxiNYC is larger than TaxiBJ (41,856 vs. 20,016), the amount of data of TaxiNYC is twice that of TaxiBJ, which significantly facilitates the pre-training. STREED-Net and STAR have impressive performance on TaxiBJ against other baselines due to the simple single-branch design. However, such simple architecture performs worse than ours in a larger dataset TaxiNYC (1.88 RMSE higher than our ST-TSNet) as there are rich spatial-temporal information that a single-branch structure can not extract effectively. Although STREED-Net and PredCNN both introduce cascading hierarchical structure in their backbone, STREED-Net has better performance than PredCNN. The reason is that STREED-Net additionally introduces channel and spatial attention mechanisms to dynamically refine the learned features to generate predictions. Nevertheless, the cascading hierarchical structure still suffers from short-range bias as it only allows distant snapshots to interact at higher layers. ST-ResNet, STAR, and PredCNN introduce a 2D convolutional layer, and MST3D, ST-3DNet, and 3D-CloST employ 3D convolution. The 3D convolutional layer is better than the 2D counterparts as it can additionally capture temporal features, while 2D convolutions are restricted to only capture spatial features. However, they all suffer from short-range bias due to the small receptive field of convolution. Moreover, they do not introduce the skip connection and any additional pre-training strategies, resulting in inferior performance.

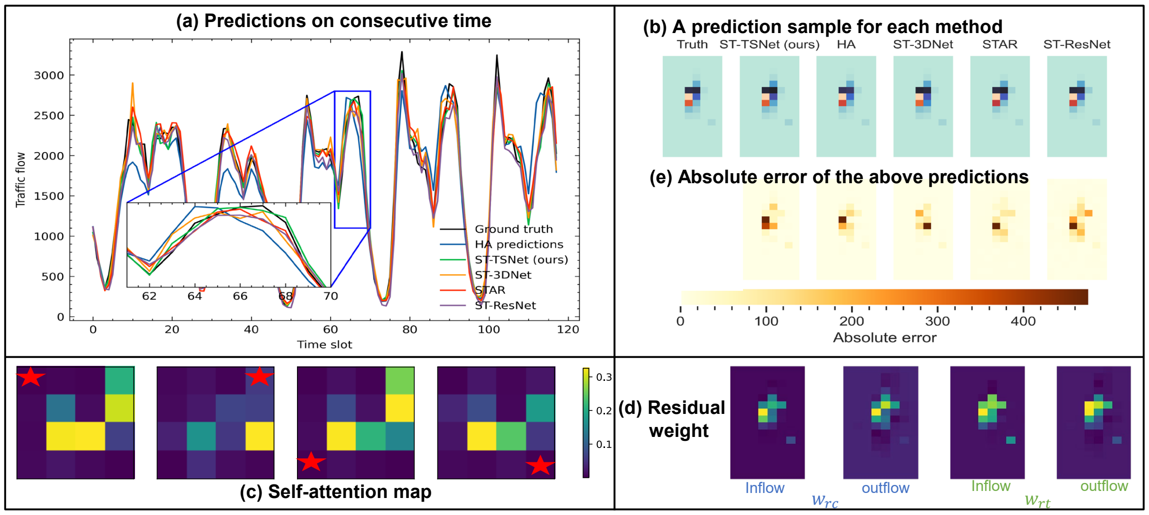

Figure 2: Qualitative analysis of our methods. (a) comparing the predicted results of each method at different time slots. (b) visualizing a prediction sample for each method and (e) showing the absolute errors of these predictions. (c) illustrating the self-attention scores of four corner patches (the pentagram-marked) for other patches and revealing that they attend to remote patches (brighter color) for long-range spatial dependencies. (d) visualizing the inflow and outflow weight of the two residual components in the fusion layer; high-flow regions usually have a higher weight.

4.4 Qualitative Analysis

We offer four intuitive visualizations of proposed methods to explain their behaviors in Fig.2. Fig.2 (a) compares the predictions of each method at different time intervals. The magnified subplot reveals that our method has better accuracy in predicting the peak. Fig.2 (b) spatially visualizes a prediction sample of each method, and Fig.2 (e) displays the absolute errors of these predictions, demonstrating that our ST-TSNet has lower prediction errors than baselines. Fig.2 (c) shows the self-attention map for four reference patches. The visualizations are produced by attention scores computed via query-key product in the ViT. We use the pentagram-marked regions as query, and show which patch (region) they attend to. The four corner patches usually attend to remote regions (brighter color meaning higher attention scores) while caring less about their neighbors. The reason is that the short-range features are perfectly captured and encoded into tokens by Pre-Conv Block, resulting in the ViT focusing more on the long-range features. Fig.2 (d) visualizes the weights of inflow and outflow of two residual components. Combining the ground truth in Fig.2 (c), we observe that although the weights vary in different regions and differ from inflow to outflow, they tend to concentrate on the regions with higher traffic flow. The reason is that these regions show a more regular time series, having more similar patterns in residual components.

4.5 Ablation Study

To verify the effectiveness of proposed methods, we design a list of variants by appending modules step by step and comparing them on TaxiBJ and TaxiNYC. The basic variant is Vision transformer (ViT). We separately append skip connection (SC), Pre-Conv Block (PC), and stochastic argumentation pre-training (SA) to ViT to construct other variants. We further consider the external factors on our ST-TSNet (ST-TSNet (w Ext)). We use an external module (two-layer multilayer perceptron) to model the external features according to [Zhang *et al.*, 2016a]. The external data is transformed and added together with the main output to yield prediction.

The results in Table 2 show that: 1) the full version of the our methods (i.e., ST-TSNet (w Ext)) achieves the best performance. 2) Adding each module step by step will progressively improve the performance. It suggests that each method is an indispensable component for our ST-TSNet.

We additionally study how connection strategy influences the skip connection by introducing a new residual component: the Pre-Conv Block output Y_{conv}. We research two connection strategies: additionally and solely connect Y_{conv} to the fusion layer. Results show that the two strategies degrade performance (1.66 and 1.35 RMSE degradation), suggesting that the Y_{conv} is harmful for prediction. Because the convolutional operations in Pre-Conv Block disrupt the semantic information in historical data (e.g., traffic distributions), resulting in the Y_{conv} and predicted target have different distributions. In contrast, the historical records (x_{trend} and x_{close}) and the predicted target are collected from the same distribution and temporally correlated. Thus the historical records share similar patterns with the predicted target that can directly contribute to the prediction, while the Y_{conv} confuses the model.

5 Conclusion

In this paper, we present a novel traffic prediction framework, spatial-temporal Transformer Network with Self-supervised Learning (ST-TSNet) for learning spatial-temporal features. ST-TSNet is equipped with Pre-Conv Block and ViT to capture local and spatial dependencies. In addition, we observe the similarity in traffic flow data, which enables us to take advantage of the historical data as the base prediction for the future. Finally, we propose a pretext task named stochastic argumentation to enable models to further explore spatial-temporal representations under limited data. Experiments on two datasets demonstrate the superiority of our proposed methods.

References

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proc. of NeurIPS*, 2020.

[Chen *et al.*, 2018] Cen Chen, K. Li, S. Teo, Guizi Chen, Xiaofeng Zou, Xulei Yang, R. Vijay, Jiashi Feng, and Zeng Zeng. Exploiting spatio-temporal correlations with multiple 3d convolutional neural networks for citywide vehicle flow prediction. *2018 IEEE International Conference on Data Mining (ICDM)*, 2018.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of ACL*, 2019.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of ICLR*, 2021.

[Fang *et al.*, 2019] Shen Fang, Qi Zhang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Gstnet: Global spatial-temporal network for traffic flow prediction. In *Proc. of IJCAI*, 2019.

[Fiorini *et al.*, 2020] Stefano Fiorini, Giorgio Pilotti, M. Ciavotta, and A. Maurino. 3d-clost: A cnn-lstm approach for mobility dynamics prediction in smart cities. *2020 IEEE International Conference on Big Data (Big Data)*, 2020.

[Fiorini *et al.*, 2021] Stefano Fiorini, Michele Ciavotta, and Andrea Maurino. Listening to the city, attentively: A spatio-temporal attention boosted autoencoder for the short-term flow prediction problem. *ArXiv preprint*, 2021.

[Gidaris *et al.*, 2018] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Proc. of ICLR*, 2018.

[Guo *et al.*, 2019] S. Guo, Youfang Lin, Shijie Li, Zhaoming Chen, and Huaiyu Wan. Deep spatial–temporal 3d convolutional neural networks for traffic data forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[Hassani *et al.*, 2021] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *ArXiv*, abs/2104.05704, 2021.

[hua Guo *et al.*, 2014] Jian hua Guo, Wei Huang, and Billy M. Williams. Adaptive kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. *Transportation Research Part C-emerging Technologies*, 2014.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.

[Lv *et al.*, 2015] Yisheng Lv, Y. Duan, Wenwen Kang, Zhengxi Li, and F. Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 2015.

[Misra and van der Maaten, 2020] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proc. of CVPR*, 2020.

[Noroozi and Favaro, 2016] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proc. of ECCV*, 2016.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proc. of CVPR*, 2016.

[Salakhutdinov, 2014] Ruslan Salakhutdinov. Deep learning. In *Proc. of KDD*, 2014.

[Wang and Su, 2019] Hongnian Wang and Han Su. Star: A concise deep learning framework for citywide human mobility prediction. *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, 2019.

[Williams and Hoel, 2003] Billy M. Williams and Lester A. Hoel. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of Transportation Engineering-asce*, 2003.

[Xu *et al.*, 2018] Ziru Xu, Yunbo Wang, Mingsheng Long, and Jianmin Wang. Predcnn: Predictive learning with cascade convolutions. In *Proc. of IJCAI*, 2018.

[Zhang *et al.*, 2016a] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. Dnn-based prediction model for spatio-temporal data. *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016.

[Zhang *et al.*, 2016b] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Proc. of ECCV*, 2016.

[Zhang *et al.*, 2017] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proc. of AAAI*, 2017.

[Zhang *et al.*, 2018] Junbo Zhang, Y. Zheng, Dekang Qi, R. Li, Xiuwen Yi, and Tianrui Li. Predicting citywide crowd flows using deep spatio-temporal residual networks. *Artif. Intell.*, 2018.

[Zhao *et al.*, 2017] Z. Zhao, Weihai Chen, Xingming Wu, Peter C. Y. Chen, and Jingmeng Liu. Lstm network: a deep learning approach for short-term traffic forecast. *Iet Intelligent Transport Systems*, 2017.