ICIC Express Letters
Volume **19**, Number **7**, July **2025**

ICIC International ©2025 ISSN 1881-803X
pp. **795–802**

DUAL-DIMENSIONAL DEPENDENCY FUSION TRANSFORMER FOR LONG-TERM SPATIOTEMPORAL FLOW PREDICTION

Younghwi Kim[1], Jaehoon Lee[1], Seungwoo Lee[1] and Sunghyun Sim[2,*]

[1]Department of Advanced Defense Engineering
[2]School of AI Convergence
Changwon National University
20 Changwondaehak-ro, Ui-chang Gu, Changwon 51140, Korea
{ dudgnl6032; younglks000 }@naver.com; dynamic97312@gmail.com
*Corresponding author: ssh@changwon.ac.kr

Received October 2024; accepted January 2025

Abstract. *With the recent surge in the collection of various real-time location sensor data, there has been an increasing number of cases where this data is utilized from a spatiotemporal prediction perspective. Spatiotemporal prediction problems are generally divided into short-term spatiotemporal flow prediction and long-term spatiotemporal flow prediction. Particularly, long-term spatiotemporal flow prediction is crucial in various applications, including traffic management, weather forecasting, and environmental monitoring. Despite advancements in deep learning improving long-term spatiotemporal prediction performance with Convolution Neural Network-based models, these models capture spatial features well but face limitations in capturing temporal features. Recently, Vision Transformers have been proposed to overcome these limitations, but they still face challenges in simultaneously capturing spatial and temporal features. In this study, we propose a Dual-dimensional Dependency Fusion (DFusion) Transformer structure designed to simultaneously capture and learn the temporal and spatial features embedded in spatiotemporal data. The proposed DFusion Transformer learns the complex dimensions of time and space through a DFusion Block composed of multiple individual Linear Layers. To demonstrate the superiority of the proposed network structure, we conducted experiments comparing long-term prediction performance with previous research results, showing that our approach achieved the best long-term prediction performance. In conclusion, this study offers an important new approach to long-term spatiotemporal flow prediction, which is expected to significantly contribute to improving prediction accuracy.*

Keywords: Long-term spatiotemporal flow prediction, Spatiotemporal data, Dual-dimensional Dependency Fusion Transformer, Dual-dimensional Dependency Fusion Block

1. **Introduction.** Spatiotemporal data typically refers to three-dimensional information that changes over time and space, involving space-time-variable dimensions [1]. This data is collected from various image sources and sensors on roads [2]. Especially, spatiotemporal data collected from various sensors on roads can be utilized to predict traffic flow within urban areas [3]. The problem of predicting spatiotemporal dynamics is typically divided into short-term and long-term spatiotemporal prediction. Both of these tasks require capturing the interdependencies of spatiotemporal data, making them challenging [4]. However, long-term spatiotemporal prediction is regarded as a more difficult problem compared to short-term spatiotemporal prediction [5]. Recently, many spatiotemporal prediction methods based on deep learning have been introduced [6]. Among them, two main approaches are 1) the combination of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) models [7,8] and 2) Transformer-based models [9,10]. Both approaches have contributed to improving prediction performance [10], but they have the

DOI: 10.24507/icicel.19.07.795

following limitations. 1) The CNN-RNN combination model effectively learns spatiotemporal information to improve performance; however, it sequentially learns spatial features over time, which limits its ability to fully capture the global spatiotemporal interdependencies of the data [11]. 2) Transformer-based models, such as ViT, can learn temporal and spatial features comprehensively [10], but they have limitations in considering both local temporal and spatial features and global ones simultaneously.

To address the limitations of previous research, we propose a new spatiotemporal learning module called the Dual-dimensional Dependency Fusion (DFusion) Block. DFusion Block enables the simultaneous learning of global and local spatiotemporal features by processing two-dimensional dependencies in parallel according to both the time-space and space-time orders. Additionally, we propose the DFusion Transformer model, hierarchically structured with the DFusion Block between the encoder and decoder, suitable for long-term spatiotemporal prediction. This approach enhances the performance of feature learning in both time and space, contributing to improved accuracy in long-term spatiotemporal prediction. The contributions of this study are as follows. 1) We have developed DFusion Block to simultaneously learn temporal and spatial features within spatiotemporal data. 2) Based on DFusion Block, we developed the DFusion Transformer model suitable for predicting long-term spatiotemporal flows. 3) We have demonstrated, using real-world data, that the performance of our proposed model significantly outperforms the baseline model by more than 15% in predicting spatiotemporal flow.

The remainder of this paper is organized as follows. Section 2 presents the related work to this study. The proposed model is described in Section 3. Sections 4 and 5 present the verification of the proposed model. Section 6 summarizes the study.

2. Related Work.

For long-term spatiotemporal flow prediction, prominent deep learning approaches include 1) Graph Neural Networks (GNN)-based models, 2) RNN, CNN-based models, and 3) Transformer-based models. In the realm of GNN-based models, notable examples include Spatial-Temporal Graph Ordinary Differential Equation (STG-ODE) Networks [12], Spatial-Temporal Graph Neural Controlled Differential Equation (STG-NCDE) [13], Dynamic Graph Convolutional Recurrent Network (DGCRN) [14], and Spatio-Temporal Graph Attention Network-based Markov Cluster Algorithm (MCL-STGAT) [15]. Firstly, STGODE and STG-NCDE leverage a combination of GNN and neural ODE to capture long-term spatiotemporal dependencies. Additionally, DGCRN constructs adjacency matrices at multiple time steps to capture dynamic correlations between nodes. MCL-STGAT proposes a spatiotemporal demand prediction model that integrates the Markov clustering algorithm with convolution operations. While these methods have shown promising performance in spatiotemporal prediction, the datasets used in the experiments are structured such that spatiotemporal predictions are conducted by assigning numbers to specific administrative regions rather than utilizing the complete grid of a designated area.

Among RNN, CNN-based models, ConvLSTM [17] stands out as a prominent example. It utilizes Long Short-Term Memory (LSTM) [18] to capture both short- and long-term spatiotemporal dependencies and employs convolutional layers for spatial modeling. However, it has been noted that processing spatial information sequentially over time, as in ConvLSTM, is inefficient. Recently, recurrent-free approaches such as CrevNet [20] and SimVP [19] have been introduced. CrevNet proposes a CNN-based reversible network to learn complex spatiotemporal interdependencies. Subsequently, SimVP presents an architecture composed entirely of complete CNN, comprising Encoder, Translator, Decoder, and Inception-Unet Translator, demonstrating state-of-the-art performance despite its simplicity. However, the method of stacking CNN is known to have inherent limitations in fully capturing spatiotemporal interdependencies due to local features [9].

Motion-Aware Unit (MAU) [10], Temporal Attention Unit (TAU) [16] and SwinLSTM [9] are prominent examples of Transformer-based models. MAU consists of an attention module that aggregates temporal states based on current and past spatial states, and a fusion module that combines the results of the attention module with the current spatial state to generate the final prediction. TAU decomposes temporal dependencies into intra-frame static attention and inter-frame dynamic attention to facilitate the parallelization of the temporal module. SwinLSTM replaces the convolutional structure of ConvLSTM with the shifted window mechanism and the hierarchical design of the Swin-Transformer attention module. This leads to a significant improvement in prediction accuracy compared to ConvLSTM. These attention mechanism-based methods have demonstrated superior performance compared to traditional approaches [9]. However, despite these performance improvements, the datasets used in their experiments are primarily focused on video frame prediction tasks. Furthermore, predictions are mainly conducted in the short term, leading to relatively lower performance in long-term prediction tasks. Therefore, we concentrate on the long-term prediction of spatiotemporal flow datasets that exhibit correlations between spatial regions along a defined trajectory.

3. **Dual-Dimensional Dependency Fusion Transformer.** In this section, we introduce the DFusion Transformer, designed for long-term spatiotemporal flow prediction based on the DFusion Block. Figure 1 illustrates the overall structure of the proposed DFusion Transformer.

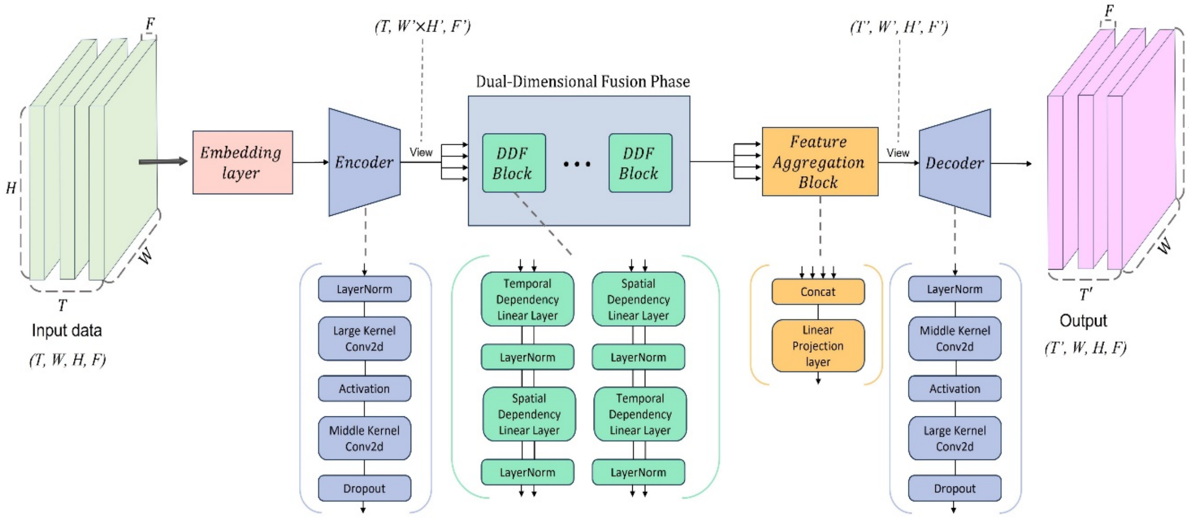
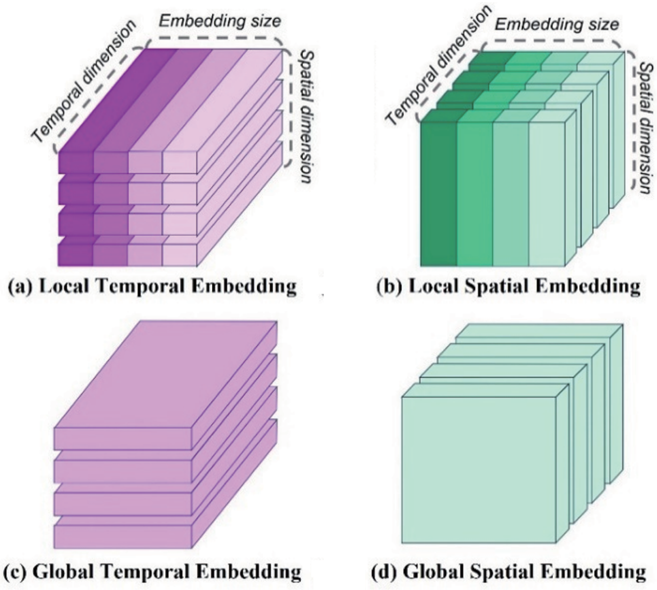


Figure 1. The overview structure of Dual-dimensional Dependency Fusion Transformer

The DFusion Transformer consists of the following four components. 1) *Embedding Layer*: This layer embeds the channel dimension of the input data into a higher-dimensional space to allow spatiotemporal data to have a high-dimensional representation within the model. This is achieved through a single Linear Layer. 2) *Encoder and Decoder*: These transform the embedded values into hidden vectors with nonlinear effects by placing an activation function between two convolutional layers. To mitigate the local calculation properties of convolution, the two convolutional layers use large kernels for wide-range and middle kernels for mid-range calculations, respectively. Additionally, the Encoder reduces the W and H dimensions while increasing the embedding dimension for computational efficiency. The Decoder is structured oppositely to the Encoder. 3) *DFusion Block*: As the core concept of DFusion Transformer, this block effectively learns the interdependencies between global and local spatiotemporal features within the high-dimensional spatiotemporal space. For local spatiotemporal interdependencies, a single Linear Layer is applied to each embedding along either the temporal or spatial dimension. For global spatiotemporal interdependencies, a single Linear Layer shared uniformly across all embeddings is applied along either the temporal or spatial dimension.

Figure 2 visually represents these concepts. Specifically, the Local Temporal Embedding applies a single Linear Layer to each embedding individually along the temporal axis, sharing weights across the spatial dimension. Similarly, the Local Spatial Embedding performs the same operation along the spatial axis, with weights shared across the temporal dimension. The Global Temporal and Spatial Embedding applies a single Linear Layer along both the temporal and spatial axes, with weights shared across all embeddings. Each of these methods allows for a detailed learning of temporal or spatial representations both globally and locally within the embedding space. These operations are alternately applied in parallel in the time-space and space-time orders, capturing their interdependencies. Consequently, the output of the DFusion Block consists of hidden features computed from four different perspectives, representing both global and local aspects in the time-space and space-time orders. 4) *Feature Aggregation Block*: This block aggregates the four types of hidden features by concatenating them along the embedding dimension. The concatenated result is then passed through a bias-free Linear Layer to form the input for the Decoder, ensuring the output reflects the contributions of each type of hidden feature.



(a) Local Temporal Embedding

(b) Local Spatial Embedding

(c) Global Temporal Embedding

(d) Global Spatial Embedding

FIGURE 2. Visualization of the application mechanism for global and local spatiotemporal dependencies in the Dual-dimensional Fusion Block

4. **Experiment Setting.** We conducted an experiment to compare the long-term spatiotemporal flow prediction accuracy of the proposed models. For this comparative experiment, we utilize two versions of the publicly available TaxiBJ dataset. This dataset, first released by [20], represents taxi trajectories in Beijing, China. Specifically, the studied area is divided into a 32×32 grid, with taxi volume measured every 30 minutes for each grid cell, resulting in a traffic flow map with a single channel. According to the data collection period, this dataset includes P1 (July 1, 2013 – October 31, 2013) and P2 (February 1, 2014 – June 30, 2014), with each part divided into a training set and a test set in an $8 : 2$ ratio. An input sequence of 30 steps is used for training, followed by predictions of taxi volume 30-step ahead (15 hours), 180-step ahead (60 hours), 360-step ahead (120 hours), and 720-step ahead (360 hours) into the future. Detailed descriptions of the training data are provided in Table 1. Additionally, Figure 4 provides examples of input data for individual sequences and corresponding output data. Ultimately, the input data

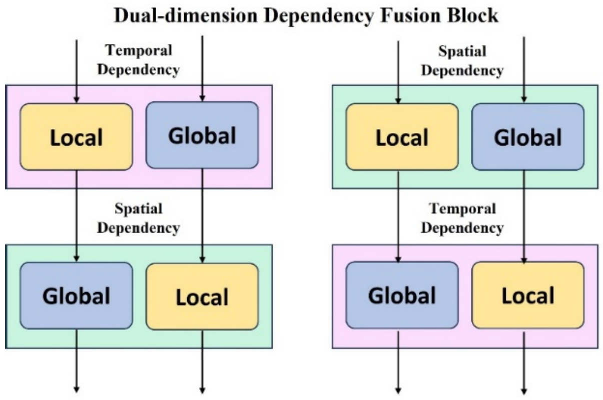


Figure 3. Structure of applying local and global computations in parallel computation of temporal-spatial and spatial-temporal interdependencies

Table 1. Detailed information about TaxiBJ (P1, P2)

Dataset	Num. train data	Num. test data	Time range	Width size	Height size	Num. features	Input sequence	Output sequence
TaxiBJ (P1)	2448	612	2013/07/01 ∼ 2013/10/31 (30 minutes)	32	32	1	30	30, 180, 360, 720
TaxiBJ (P2)	2847	712	2014/02/01 ∼ 2014/06/30 (30 minutes)	32	32	1	30	30, 180, 360, 720

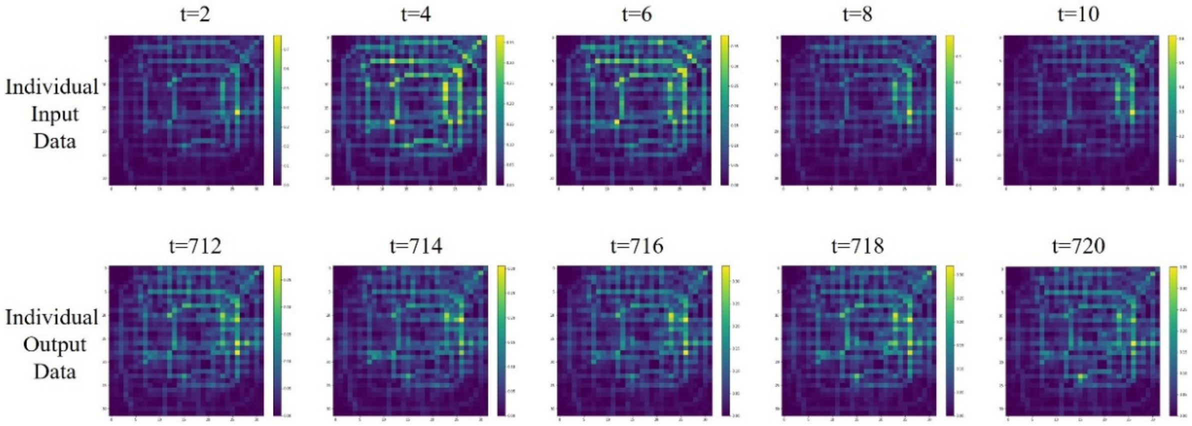


Figure 4. Visual examples of individual input data and individual output data

consolidates 30 individual sequence inputs from Figure 4 into a single dataset. Furthermore, the output data generates 30, 180, 360, and 720 individual sequences respectively, considering the prediction time points.

For the comparative experiment, we selected representative models from RNN-based, CNN-based, and Transformer-based categories, commonly used for spatiotemporal flow prediction, as baselines. Ultimately, we conducted comparative experiments between the RNN-based model ConvLSTM, the CNN-based model SimVP, and the Transformer-based models TAU and MAU. To compare prediction performance, we used four metrics: 1) Mean Squared Error (MSE), 2) Mean Absolute Error (MAE), 3) Peak Signal-to-Noise Ratio (PSNR), and 4) Structural Similarity Index Measure (SSIM). These metrics allow us to evaluate prediction accuracy from various perspectives. The formulas representing these metrics are as follows, with lower values of MSE and MAE indicating better prediction accuracy, while higher values of PSNR and SSIM signify improved prediction performance.

$$MSE = \frac{\sum_{N,W,H} \left(y_{n,w,h} - \hat{y}_{n,w,h}\right)^2}{N \cdot W \cdot H} \tag{1}$$

$$MAE = \frac{\sum_{N,W,H} \left|y_{n,w,h} - \hat{y}_{n,w,h}\right|}{N \cdot W \cdot H} \tag{2}$$

$$PSNR = 10 \cdot \log_{10}\left(\frac{R^2}{MSE}\right) \tag{3}$$

$$SSIM = \frac{1}{N}\sum_{N} \frac{\left(2\mu_{y_n}\mu_{\hat{y}_n} + C_1\right)\left(2\sigma_{y_n\hat{y}_n} + C_2\right)}{\left(\mu_{y_n}^2 + \mu_{\hat{y}_n}^2 + C_1\right)\left(\sigma_{y_n}^2 + \sigma_{\hat{y}_n}^2 + C_2\right)} \tag{4}$$

Here, N denotes the prediction range (number of grids), while W and H represent the grid resolution, and R equals 255. Additionally, μ_y, $\mu_{\hat{y}}$, σ_y, $\sigma_{\hat{y}}$ and $\sigma_{y\hat{y}}$ represent the local mean, standard deviation, and cross-covariance of the grid, respectively. The constants C_1 and C_2 are small values used to prevent instability in the calculations, defined as $C_1 = 0.01 \times R$ and $C_2 = 0.03 \times R$.

5. **Experiment Results.** Tables 2 and 3 present the comparison results for 30-step ahead, 180-step ahead, 360-step ahead, and 720-step ahead predictions on the TaxiBJ (P1) and TaxiBJ (P2) datasets, respectively. For each dataset, the best result for each prediction term is highlighted in red, while the second-best result is underlined in blue. According to the results in Tables 2 and 3, the proposed DFusion Transformer demonstrated the best average prediction performance, followed by SimVP, MAU, TAU and ConvLSTM in descending order of prediction accuracy. The proposed method showed a 27.36% improvement in prediction performance compared to ConvLSTM, a 12.60% improvement over MAU, and a 6.95% improvement over SimVP.

TABLE 2. The comparison results of long-term spatiotemporal flow prediction performance for 30-step ahead and 180-step ahead

Dataset	Method	30				180			
		MSE	*MAE*	*PSNR*	*SSIM*	*MSE*	*MAE*	*PSNR*	*SSIM*
TaxiBJ (P1)	ConvLSTM	11.98	1.82	34.044	0.877	11.44	1.95	33.533	0.867
	SimVP	7.39	1.41	35.220	0.918	10.15	**1.75**	**34.086**	**0.890**
	MAU	7.35	1.37	**35.501**	0.921	12.48	2.00	33.307	0.868
	TAU	7.90	1.38	35.372	0.914	**9.59**	1.81	34.052	0.888
	Ours	**7.02**	**1.36**	35.432	**0.924**	9.80	1.76	34.058	0.889
TaxiBJ (P2)	ConvLSTM	7.49	1.59	34.451	0.923	10.24	1.91	33.459	0.899
	SimVP	3.86	**0.98**	**37.061**	**0.966**	9.20	1.77	33.847	0.912
	MAU	5.00	1.23	35.808	0.950	10.13	1.86	33.754	0.904
	TAU	5.14	1.31	35.707	0.946	9.13	1.67	34.129	0.914
	Ours	**3.56**	1.06	36.565	0.960	**8.50**	**1.65**	**34.212**	**0.915**

6. **Conclusions.** In this study, we propose a novel network architecture suitable for long-term spatiotemporal flow prediction. Previous research has explored various approaches to learn spatial and temporal features within the spatiotemporal domain, but these approaches have limitations in learning both global and local spatiotemporal features contained in the data. To address these limitations, our proposed model introduces local temporal embeddings, spatial embeddings, global temporal embeddings, and spatial embeddings, and proposes the DFusion Block module to learn interdependencies among these embeddings.

TABLE 3. The comparison results of long-term spatiotemporal flow prediction performance for 360-step ahead and 720-step ahead

Dataset	Method	360				720			
		MSE	*MAE*	*PSNR*	*SSIM*	*MSE*	*MAE*	*PSNR*	*SSIM*
TaxiBJ (P1)	ConvLSTM	11.26	1.97	33.412	0.871	12.40	2.12	33.102	0.856
	SimVP	10.48	1.82	33.806	0.881	12.51	2.01	33.402	0.861
	MAU	12.93	2.04	33.236	0.864	12.74	**1.87**	**33.819**	**0.874**
	TAU	10.30	1.79	33.872	**0.887**	12.65	2.06	33.321	0.857
	Ours	**10.14**	**1.78**	**33.912**	0.884	**12.14**	2.00	33.424	0.862
TaxiBJ (P2)	ConvLSTM	12.81	2.16	32.893	0.877	11.31	2.05	33.102	0.889
	SimVP	10.51	1.88	33.544	0.902	11.01	2.03	33.189	0.896
	MAU	10.09	1.86	33.614	0.903	**9.00**	**1.73**	**33.841**	**0.916**
	TAU	10.19	1.91	33.644	0.904	9.76	1.77	33.641	0.909
	Ours	**9.92**	**1.82**	**33.657**	**0.905**	9.41	1.81	33.692	0.910

Finally, we perform long-term spatiotemporal flow prediction using the DFusion Transformer designed with the DFusion Block module. Comparative experiments to validate the performance of the proposed model showed a 15.63% improvement over previous studies. We expect that our proposed model can contribute to various real-world applications that utilize long-term spatiotemporal flow prediction. Given that various external factors significantly influence spatiotemporal data, we plan to develop our model to incorporate these factors in future research. Additionally, we aim to conduct further experiments using a wider range of spatiotemporal flow datasets and long-term prediction scenarios. We hope that our research will contribute to various real-world domains that utilize long-term spatiotemporal flow prediction and facilitate further advancements in this area.

Acknowledgment. This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201791, 50%) and was supported by the 5th Educational Training Program for the Shipping, Port and Logistics from the Ministry of Oceans and Fisheries (50%).

REFERENCES

[1] E. Koutsaki, G. Vardakis and N. Papadakis, Spatiotemporal data mining problems and methods, *Analytics*, vol.2, no.2, pp.485-508, 2023.

[2] L. Jiang, T. Zhang, Q. Zuo, C. Tian, G. P. Chan and W. K. V. Chan, A deep learning framework for traffic data imputation considering spatiotemporal dependencies, *2022 IEEE 7th International Conference on Intelligent Transportation Engineering (ICITE)*, pp.14-19, 2022.

[3] M. Jin, Q. Wen, Y. Liang, C. Zhang, S. Xue, X. Wang, J. Zhang, Y. Wang, H. Chen, X. Li, S. Pan and V. S. Tseng, Large models for time series and spatio-temporal data: A survey and outlook, *arXiv Preprint*, arXiv: 2310.10196, 2023.

[4] Z. Lin, M. Li, Z. Zheng, Y. Cheng and C. Yuan, Self-attention convlstm for spatiotemporal prediction, *Proc. of the AAAI Conference on Artificial Intelligence*, vol.34, no.7, pp.11531-11538, 2020.

[5] D. Iskandaryan, F. Ramos and S. Trilles, Spatiotemporal prediction of nitrogen dioxide based on graph neural networks, in *Advances and New Trends in Environmental Informatics. ENVIROINFO 2022. Progress in IS*, V. Wohlgemuth, S. Naumann, G. Behrens, H. K. Arndt and M. Höb (eds.), Cham, Springer International Publishing, 2022.

[6] G. Jin, Y. Liang, Y. Fang, Z. Shao, J. Huang, J. Zhang and Y. Zheng, Spatio-temporal graph neural networks for predictive learning in urban computing: A survey, *IEEE Transactions on Knowledge and Data Engineering*, vol.36, pp.5388-5408, 2023.

[7] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong and W. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, *Advances in Neural Information Processing Systems*, vol.28, 2015.

[8] N. Ballas, L. Yao, C. Pal and A. Courville, Delving deeper into convolutional networks for learning video representations, *arXiv Preprint*, arXiv: 1511.06432, 2015.

[9] S. Tang, C. Li, P. Zhang and R. N. Tang, SwinLSTM: Improving spatiotemporal prediction accuracy using Swin Transformer and LSTM, *Proc. of IEEE/CVF International Conference on Computer Vision*, pp.13470-13479, 2023.

[10] Z. Chang, X. Zhang, S. Wang, S. Ma, Y. Ye, X. Xiang and W. Gao, MAU: A motion-aware unit for video prediction and beyond, *Advances in Neural Information Processing Systems*, vol.34, pp.26950-26962, 2021.

[11] V. L. Guen and N. Thome, Disentangling physical dynamics from unknown factors for unsupervised video prediction, *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11474-11484, 2020.

[12] Z. Fang, Q. Long, G. Song and K. Xie, Spatial-temporal graph ODE networks for traffic flow forecasting, *Proc. of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp.364-373, 2021.

[13] J. Choi, H. Choi, J. Hwang and N. Park, Graph neural controlled differential equations for traffic forecasting, *Proc. of AAAI Conference on Artificial Intelligence*, vol.36, no.6, pp.6367-6374, 2022.

[14] J. Li, J. Feng, H. Yan, G. Jin, F. Yang, F. Sun, D. Jin and Y. Li, Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution, *ACM Transactions on Knowledge Discovery from Data*, vol.17, no.1, pp.1-21, 2023.

[15] T. Zhang, Y. Wang and Z. Wei, MCL-STGAT: Taxi demand forecasting using spatio-temporal graph attention network with Markov cluster algorithm, *International Journal of Innovative Computing, Information and Control*, vol.19, no.4, pp.1251-1264, 2023.

[16] C. Tan, Z. Gao, L. Wu, Y. Xu, J. Xia, S. Li and S. Z. Li, Temporal attention unit: Towards efficient spatiotemporal predictive learning, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.18770-18782, 2023.

[17] N. Srivastava, E. Mansimov and R. Salakhudinov, Unsupervised learning of video representations using LSTMs, *International Conference on Machine Learning*, vol.37, pp.843-852, 2015.

[18] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation*, vol.9, no.8, pp.1735-1780, 1997.

[19] Z. Gao, C. Tan, L. Wu and S. Z. Li, SimVP: Simpler yet better video prediction, *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.3170-3180, 2022.

[20] J. Zhang, Y. Zheng and D. Qi, Deep spatio-temporal residual networks for citywide crowd flows prediction, *Proc. of AAAI Conference on Artificial Intelligence*, vol.31, no.1, 2017.