Hindawi
Journal of Advanced Transportation
Volume 2024, Article ID 4349402, 17 pages
https://doi.org/10.1155/2024/4349402

Research Article

Air Traffic Flow Prediction with Spatiotemporal Knowledge Distillation Network

Zhiqi Shen,[1] **Kaiquan Cai**,[1] **Quan Fang**,[2] **and Xiaoyan Luo**[3]

[1]*School of Electronic Information Engineering, Beihang University, Beijing 100191, China*
[2]*School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[3]*School of Astronautics, Beihang University, Beijing 100191, China*

Correspondence should be addressed to Kaiquan Cai; caikq@buaa.edu.cn

Received 28 August 2023; Revised 1 April 2024; Accepted 26 April 2024; Published 15 May 2024

Academic Editor: Yuchuan Du

Accurate air traffic flow prediction assists controllers formulate control strategies in advance and alleviate air traffic congestion, which is important to flight safety. While existing works have made significant efforts in exploring the high dynamics and heterogeneous interactions of historical air traffic flow, two key challenges still remain. (1) The transfer patterns of air traffic are intricate, subject to numerous constraints and limitations such as controllers, flight regulations, and other regulatory factors. Relying solely on mining historical traffic evolution patterns makes it difficult to accurately predict the constrained air traffic flow. (2) Weather conditions exert a substantial influence on air traffic, making it exceptionally difficult to simulate the impact of external factors (such as thunderstorms) on the evolution of air traffic flow patterns. To address these two challenges, we propose a Spatiotemporal Knowledge Distillation Network (ST-KDN) for air traffic flow prediction. Firstly, recognizing the inherent future insights embedded within flight plans, we develop a "teacher-student" distillation model. This model leverages the prior knowledge of upstream-downstream migration patterns and future air traffic trends inherent in flight plans. Subsequently, to model the influence of external factors and predict air traffic flow disturbed by thunderstorm weather, we propose a student network based on the "parallel-fusion" structure. Finally, employing a feature-based knowledge distillation approach to integrate prior knowledge from flight plans and extract meteorological features, our method can accurately capture complex and constrained spatiotemporal dependencies in air traffic and explicitly model the impact of weather on air traffic flow. Experimental results on real-world flight data demonstrate that our method can achieve better prediction performance than other state-of-the-art comparison methods, and the advantages of the proposed method are particularly prominent in modeling the complicated transfer pattern of air traffic and inferring nonrecurrent flow patterns.

1. Introduction

With the rapid development of civil aviation industry, the number of aircraft has greatly increased, and thus air congestion and flight delays occur frequently [1–3]. External factors such as thunderstorm weather have aggravated the contradiction between the air traffic demand and the limited capacity of air traffic management (ATM) system. Air traffic flow management (ATFM), recognized as a widely implemented and effective strategy, plays a pivotal role in ensuring efficient and safe air transportation operations [4]. Air traffic flow prediction, as the key part of the ATFM system, helps the controllers to formulate control strategies in advance, thereby alleviating air traffic congestion [5, 6].

Researchers have already proposed many methods to predict air traffic flow. Early researchers mainly used dynamic simulation algorithms; however, these methods have high computational complexity, especially when the number of aircraft is increasing greatly [7, 8]. Recently, deep learning methods have received considerable attention. Some researchers used convolutional neural networks (CNNs) and long short-term memory (LSTM) [9] to model temporal and spatial correlations. In contrast, numerous researchers in road traffic used graph convolution network (GCN) to capture the topological features of traffic networks, such as

spatiotemporal graph convolution network (STGCN) [10], attention-based spatiotemporal graph convolution network (ASTGCN) [11], and adaptive graph convolutional recurrent network (AGCRN) [12]. In response to the complexities inherent in dynamic and time-delayed traffic data, a novel propagation delay-aware dynamic long-range transformer (PDFormer) model leveraging a spatiotemporal self-attention mechanism has recently been introduced [13]. Recognizing the impact of spatiotemporal heterogeneity on traffic prediction, Ji et al. put forward a self-supervised learning framework [14], integrating an adaptive heterogeneity-aware enhancement scheme into the spatiotemporal graph structure to mitigate noise disturbances.

Despite the promising performance of introducing GCN in traffic field, we argue that there are several important aspects that previous methods have overlooked.

(i) Firstly, air traffic flow has complicated and constrained transfer pattern. To ensure safety, flights must not only use predefined routes as guidance but also follow the instruction of air traffic controllers [15]. Existing methods mostly learn the spatiotemporal correlations of flow patterns among different nodes over different time intervals from historical traffic data to infer future traffic [10, 11], as illustrated in Figure 1(a). While numerous researchers have made significant efforts to learn the complex and constrained spatiotemporal dependencies in air traffic [12, 13], they still cannot achieve satisfactory performance in air traffic flow prediction. It is noteworthy that the flight plan in air traffic management contains some predefined rule constraints and provides effective prior knowledge of future regular evolution patterns. However, they have not been fully utilized. Therefore, we try to combine the valuable prior knowledge in the currently underutilized flight plan information to develop a more efficient and effective solution. As depicted in Figure 1(b), flight plans offer inherent insights into future air traffic dynamics. They provide information of how air traffic flow transits from each node to another, thereby implying the dependency of upstream and downstream flows. The dependency embeds future knowledge of how downstream traffic is caused by upstream traffic, thus aiding in more accurate inference of future traffic patterns.

(ii) Secondly, prevailing methodologies often overlook the substantial influence of external variables, such as weather, on the dynamic evolution of air traffic flow [14, 15]. Adverse weather conditions wield considerable impact, such as localized thunderstorms not only disrupting regional air traffic but also spreading into global airspace. Although some researchers embed weather conditions and accidents into the spatiotemporal learning framework to predict the nonrecurrent road traffic flow [16–18], they cannot be directly applied to modeling the effects of weather on air traffic flow. Thus, how to effectively model the impact of weather on air traffic flow patterns is still unresolved.

Reference [19] proposes a temporal attention-aware dual-graph convolution network (TAaDGCN) to predict air traffic flow under regular conditions. To capture the spatial dependencies, a dual-graph convolution module and spatial embedding (SE) block are designed. To capture the temporal dependencies of historical traffic, attention mechanisms are utilized. Through the spatiotemporal modeling module, the TAaDGCN method has learned the spatiotemporal evolution patterns of historical air traffic flow under regular conditions. Compared with [19], we propose a Spatiotemporal Knowledge Distillation Network (ST-KDN) to predict air traffic flow under the influence of other factors such as thunderstorms. Differing from most existing methods that solely learn spatiotemporal dependencies from historical traffic data, we fully exploit the prior knowledge of future insights embedded within flight plans, including predefined rule constraints, to more accurately predict future air traffic flow. Specifically, considering that flight plan information provides inherent insights into future air traffic dynamics and reflects regular flow evolution patterns, we design a teacher network that incorporates flight plan data. Then, to comprehensively capture the effects of adverse weather, including thunderstorms, on air traffic flow, we design a student network structured upon a "parallel-fusion" architecture. This network comprises two distinct components: one is dedicated to learning regular air traffic flow evolution patterns and the other focuses on weather variation characteristics. Subsequently, a feature fusion module is crafted to integrate the features of both regular air traffic flow and weather. By amalgamating prior knowledge embedded within flight plans and incorporating meteorological features, our method can explicitly capture complex spatiotemporal dependencies of air traffic flow. Our main contributions are summarized as follows:

(i) We identify the unique characteristics of air traffic flow evolution and propose a spatiotemporal knowledge distillation network specifically for air traffic flow prediction. It effectively leverages the inherent capability of flight plans to provide insights into future air traffic dynamics, thereby enhancing prediction accuracy, especially in long-term prediction.

(ii) We consider the impact of external thunderstorm weather on spatiotemporal modeling and design a student network based on the "parallel-fusion" structure to explicitly model the impact of weather on air traffic flow, making predictions more robust.

(iii) We conduct extensive experiments with real-world flight data and meteorological radar echo data. The results suggest that the proposed method outperforms state-of-the-art approaches and is especially superior in long-term prediction of nonrecurrent flow patterns affected by weather.

The rest of the paper is organized as follows. Section 2 reviews relevant research about air traffic flow prediction and knowledge distillation. Section 3 gives the problem statement and some preliminaries. In Section 4, the

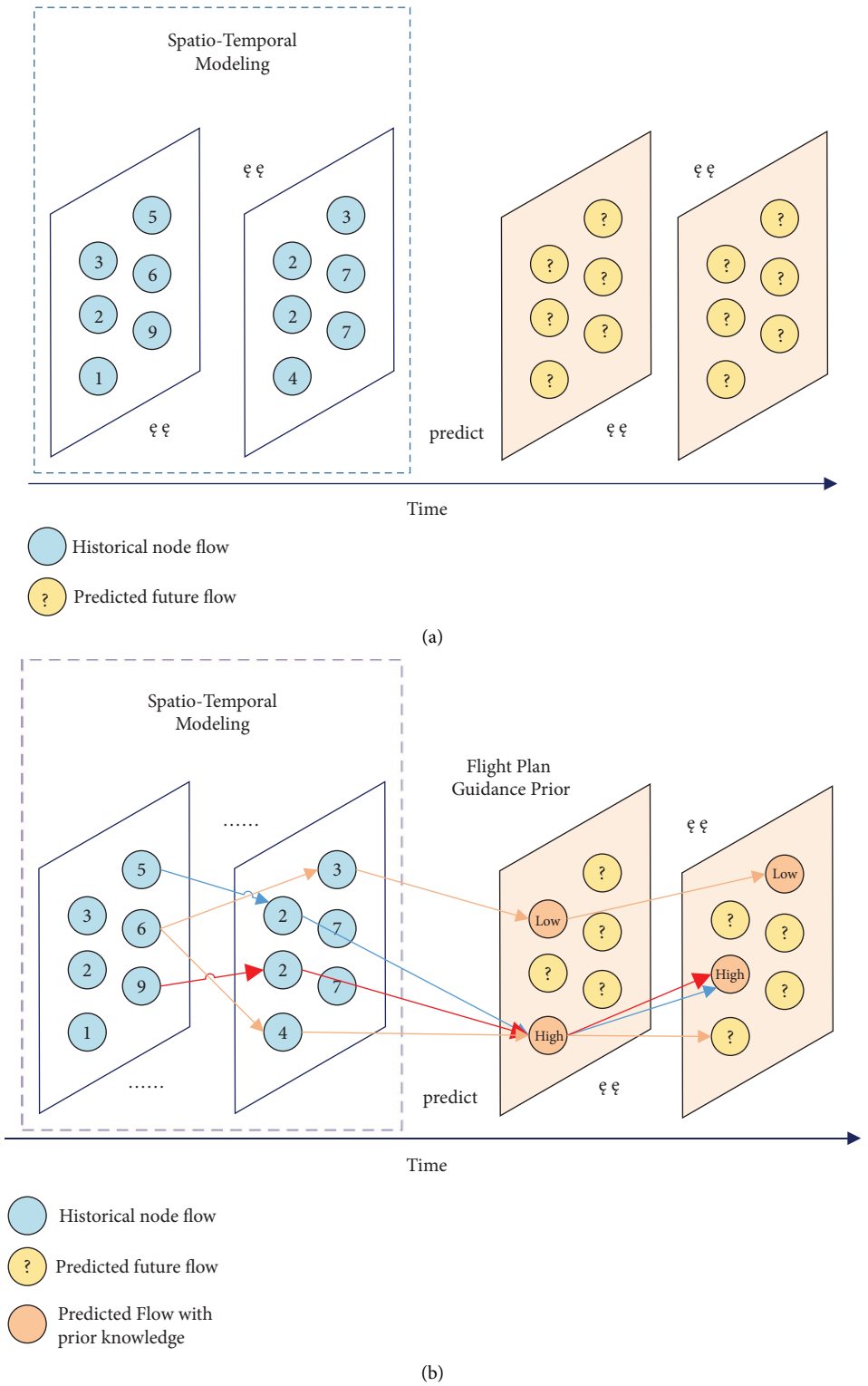

Figure 1: The flowchart about how flight plan information could aid in predicting air traffic flows. (a) When only flow observations are available. (b) When the flight plan is utilized.

framework of the proposed ST-KDN network is presented in detail. In Section 5, we evaluate the predictive performance of the proposed ST-KDN network with real-world flight data and meteorological radar echo data, including model comparisons, variant comparisons, and case study. In addition, we conclude the paper in Section 6.

2. Related Work

In this section, we offer a thorough overview of contemporary research in spatiotemporal prediction, focusing on three key areas. Firstly, we outline advancements and challenges in air traffic flow prediction. Secondly, given that air traffic flow constitutes typical spatiotemporal data, the mining of such data significantly impacts prediction performance. Consequently, we provide a summary of research on spatiotemporal semantic understanding. Lastly, we summarize the current status and applications of knowledge distillation methods.

2.1. Air Traffic Flow Prediction. In recent years, air traffic flow prediction has attracted widespread attention from researchers all over the world [18, 20–22]. In this subsection, we provide an overview of some representative traffic prediction methods, categorizing them into statistical-based methods, traditional machine learning methods, and deep learning methods.

In statistical-based methods, dynamic simulation algorithms and time series prediction algorithms have been focused on. Early attempts use dynamical simulation algorithms to model air traffic problems [7], for example, the real-time flight plan data in the EuroCat-X system are utilized to predict positioning points, airports, routes, sectors, etc. [7, 8]. However, these methods require complex system programming and high computational complexity. Since the flight flow of a region for a certain period is affected by the flight flow of first several hours, it could utilize the flow data of the first few hours to predict the flow of the subsequent period. Therefore, some researchers model it as a time series problem. Autoregressive Moving Average (ARMA) [23] is a fundamental time series forecasting method, with its variant being Autoregressive Integrated Moving Average (ARIMA) [24]. In References [25, 26], ARIMA is employed to model the temporal correlations of air traffic flow, establishing a combination model of autoregression, differencing, and moving average by analyzing the autocorrelation and partial autocorrelation properties of the sequences. It requires minimal domain knowledge and is capable of capturing both long-term and short-term trends in the data. Vector Autoregression (VAR) [27] is also widely used in time series-based traffic flow prediction, by constructing a linear relationship model between multiple time series, while considering the mutual influence of each sequence. As another extension of the ARMA method, the Seasonal Autoregressive Integrated Moving Average (SARIMA) method [28] can capture the inherent correlations in time series data, particularly suitable for modeling seasonal and random time series commonly found in traffic flow data. Although these classical time series methods can capture the temporal dependencies in time series data, they rely on strong linear and stationarity assumptions, often neglecting the spatial dependencies of neighboring regions.

With the rapid development of artificial intelligence, data-driven methods have received considerable attention [29], leading to the emergence of various traditional machine learning-based approaches, such as k-Nearest Neighbors (k-NN) [30] and support vector regression (SVR) [31, 32]. Qiu and Li proposed an air traffic flow prediction method considering wavelet neural network, which uses nonlinear wavelet to replace the nonlinear activation function in classical neural network [33]. Zhang et al. proposed an air traffic prediction model based on support vector machines to improve the real-time monitoring and control in terminal areas [32]. Zhu et al. investigated the application of Linear Conditional Gaussian (LCG) Bayesian Network (BN) models for short-term traffic flow prediction, considering both spatial-temporal features and velocity information [34]. From these preceding conventional machine learning approaches, it can be concluded that machine learning is a powerful tool in air traffic flow management (ATFM). However, the proliferation of traffic sensors in recent years along with the rapid advancement of intelligent transportation systems has led to an explosion of traffic data. Conventional machine learning methods are limited in uncovering deep, latent spatiotemporal correlations within large-scale traffic data, thereby constraining their prediction capability.

Deep learning-based methodologies are emerging as popular techniques for spatiotemporal tasks in transportation. The success of deep learning in numerous application domains, driven by the availability of big data and robust computational resources, has propelled its adoption in the field of traffic flow prediction [35–38]. Some researchers attempted to model spatiotemporal correlation in air traffic by CNN and LSTM [9, 39]. To capture the topological characteristics of traffic networks, graph convolution network (GCN) is used in road traffic network [40]. Yu et al. proposed a STGCN method, which models traffic networks as graphs and utilizes GCN to learn spatiotemporal dependencies among nodes [10]. Guo et al. proposed a novel attention-based spatiotemporal graph convolutional network (ASTGCN) to address traffic flow prediction, composed of three independent components modeling three temporal attributes of traffic flow: short-term, daily periodic, and weekly periodic dependencies [11]. Bai et al. proposed an adaptive graph convolutional recurrent network (AGCRN) to automatically capture fine-grained spatiotemporal correlations in traffic sequences [12]. Ma et al. proposed an improved long short-term memory (LSTM) network combining forward and backward LSTMs to incorporate long-term dependencies, effectively overcoming significant prediction errors [41]. Recently, a Multiview Dynamic Graph Convolutional Network (MVDGCN) has been proposed to capture diverse levels of spatiotemporal dependencies. Leveraging coupled graph convolutional networks, it dynamically learns the relationship matrix between stations, thereby capturing spatial dependencies at various levels within the traffic network [42]. Rajeh et al. proposed a deep learning-assisted method based on traffic flow dependencies and dynamics. By explicitly integrating spatiotemporal flow dependencies, traffic dynamics, and deep learning techniques, it predicts high-resolution traffic speed propagation across the network. The effective combination of physical models with deep learning methods within this framework, evolving them jointly, enhances

prediction performance [43]. To address the challenges posed by the dynamic and time-delayed nature of complex traffic data, a Propagation Delay-aware dynamic long-range transformer (PDFormer) model is proposed [13]. This model incorporates a spatial self-attention module that models local geographic neighborhoods and global semantic neighborhoods through different graph masking techniques. Additionally, a traffic delay-aware feature transformation module is devised to explicitly model time delays in the spatial information propagation. Considering the interference of spatiotemporal heterogeneity on traffic prediction, Ji et al. proposed a self-supervised learning framework [14]. An adaptive heterogeneity-aware enhancement scheme is applied to the spatiotemporal graph structure to address noise disturbances. By integrating two self-supervised learning tasks, the method enhances the capability of discerning spatial and temporal traffic heterogeneity, effectively accomplishing traffic prediction tasks. These methodologies focus on learning spatiotemporal dependencies from extensive historical data but still fail to achieve satisfactory performance in modeling rare scenarios from history and predicting long-range traffic.

2.2. Spatiotemporal Semantic Understanding. Spatiotemporal semantic understanding refers to the process of analyzing and comprehending spatiotemporal data, aiming to reveal the spatiotemporal relationships, semantic information, and patterns within the data [44–46]. Air traffic flow represents a typical form of spatiotemporal data, where the depth of exploration into spatiotemporal traffic flow data directly determines the quality of predictive performance.

In recent years, spatiotemporal semantic understanding has witnessed significant advancements across various domains. In the field of image and video understanding, Yin et al. introduced a spatiotemporal semantic understanding method based on a spatiotemporal tag library for automatic video annotation, effectively mining the complex semantic information within the tag library [47]. To address challenges posed by spatiotemporal data in dimensions, distributions, and inherent informational content, a Semantic-Aware Adaptive Knowledge Distillation Network (SAKDN) is proposed [48]. It enhances action recognition in visual sensor modalities (videos) by adaptively transferring and refining knowledge from multiple spatiotemporal data sources, effectively highlighting critical areas within complex spatiotemporal data while retaining the interrelationships of the original data. Additionally, to leverage rich spatiotemporal knowledge and generate effective supervisory signals from extensive unannotated spatiotemporal data, Liu et al. utilized multiscale temporal dependencies in videos and proposed a novel video self-supervised learning framework called Time Contrastive Graph Learning (TCGL) [49]. This framework effectively learns global contextual representations of complex spatiotemporal knowledge. Furthermore, to integrate domain-invariant representation learning and cross-modal feature fusion into a unified optimization framework, a Deep Image to Video Adaptation and Fusion Network (DIVAFN) is introduced [50]. Training action recognition classifiers demonstrate the effectiveness of this approach in learning relevant complementary knowledge.

In the realm of transportation, a series of methods for traffic identification, prediction, and planning based on spatiotemporal semantic understanding have been proposed [51–53]. Lin introduced a novel Reinforcement Learning-(RL-) based Traffic Signal Control (TSC) method named DenseLight, which employs an unbiased reward function to provide dense feedback on policy effectiveness [54]. Additionally, it utilizes a nonlocal enhanced TSC agent to predict future traffic conditions more accurately, enabling more precise traffic control. Wang et al. proposed a POI-MetaBlock network that utilizes the functionality of each region (represented by the distribution of points of interest) as metadata to further explore different traffic features within regions with different functionalities [55]. This model can be seamlessly integrated into traditional traffic flow prediction models, significantly enhancing prediction performance.

2.3. Knowledge Distillation. Knowledge distillation is a model-independent strategy that transfers the knowledge from the pretrained teacher network to guide the training of the student network. Knowledge distillation was originally proposed for model compression [56, 57]. By learning the knowledge of the large teacher network, the lightweight student network can achieve results close to or even better than the teacher network [58–60]. Kang et al. proposed a hierarchical topological distillation model for recommender systems by transforming a topology built on teacher spatial relations [61]. Dai et al. proposed a novel general instance distillation method for the object detection task, which is based on discriminable instances without considering the positive and negative distinguished by ground truth [62]. Passban et al. proposed an attention-dependent combined knowledge distillation technique, which fuses teacher-side information and takes each layer's significance into consideration [63].

In addition to model compression, due to the flexible teacher-student architectures and knowledge transfer, knowledge distillation has been applied to many other fields, such as cross-modal learning [64–66], multitask learning [67, 68], and transfer learning [69]. Thoker and Gall proposed a cross-modal knowledge distillation network to address the problem on action recognition. The network has been trained on a modality like RGB videos that can be adapted to recognize actions for another modality like sequences of 3D human poses [70]. Zhao et al. designed a novel knowledge distilling network, which considers the different distances between multiple sources and the target and the different similarities of the source samples into the target ones for multisource distilling domain adaptation [71]. Lu et al. proposed a novel knowledge distillation framework for high-dimensional search indexes, aiming to efficiently learn lightweight indexes by distilling knowledge from high-precision graph-based indexes [72]. Yang et al. introduced a Mutual Contrastive Learning (MCL) framework for online knowledge distillation, with the core idea

being the mutual interaction and transfer of contrastive distributions among a cohort of networks in an online manner [73].

To achieve knowledge distillation, early works make attempts by matching the class distribution (i.e., softmax output) [74]. The idea of class distribution-based knowledge is straightforward and easy to understand. From another perspective, the effectiveness of class distribution is similar to that of label smoothing or regularization. However, the researchers observe that utilizing output feature alone is insufficient since meaningful intermediate information may be ignored [75, 76]. Therefore, subsequent methods utilize teacher's middle layer along with output layer to distill knowledge [77]. Not only that, some researchers also use relationships between different layers as guide for student network training. The relation-based methods further explore the relationships between different layers or data samples [78, 79].

Motivated by the aforementioned techniques and considering the unique characteristics of air traffic flow, we have devised a "teacher-student" distillation framework. Unlike most methods that primarily focus on capturing the spatiotemporal relationships of traffic flow from historical data, our method adeptly leverages the prior knowledge embedded within flight plans through knowledge distillation, thus providing efficient information for future air traffic prediction. Additionally, a student network of "parallel-fusion" architecture is designed to effectively model the impact of external factors such as thunderstorms on the variation of air traffic flow. Compared to prevailing methods, our proposed method is better suited for predicting rare patterns in historical data and is particularly effective for long-term prediction.

3. Preliminaries

3.1. Problem Statement. In air traffic flow prediction, the objective is to predict future air traffic flow given historical traffic flow.

Definition 1 (Air Traffic Flow). We regard all the airspace as a graph structure, and each subregion is a node $v \in V$ of the graph, where V represents the subregion set. Each node on the network generates a flow vector $X_k = [X_{t_0,k}, X_{t_1,k}, \ldots, X_{t_j,k}, \ldots] \in R^Q$. The flow of all subregion at the t_j-time slice is represented as $X_{t_j} = [X_{t_j,1}, X_{t_j,2}, \ldots, X_{t_j,N}] \in R^N$, where N is the number of subregion and $X_{t_j,k} \in R^1$ represents the air traffic flow of the k-th subregion at the t_j-th time slice. Specifically, actual air traffic flow $X^{\text{actual}}_{t_j,k}$ can be calculated as

$$X^{\text{actual}}_{t_j,k} = \sum_{i=1}^{N_f} \sum_{j=0}^{m_i'} y,$$
$$y = \begin{cases} 1, & l_{f^a_{ij}} \in k, T_{f^a_{ij}} \in t_j, \\ 0, & \text{others}, \end{cases} \tag{1}$$

where f^a_{ij} represents j-th real trajectory point of i-th flight, $l_{f^a_{ij}}$ represents latitude and longitude of trajectory point f^a_{ij}, and $T_{f^a_{ij}}$ represents time corresponding to trajectory point f^a_{ij}.

Definition 2 (Flight Plan). Suppose there are N_f planned flight trajectories: $[F^p_1, F^p_2, \ldots, F^p_i, \ldots, F^p_{N_f}]$, and the i-th planned trajectory F^p_i can be represented as : $[f^p_{i0}, f^p_{i1}, \ldots, f^p_{im_i}]$, where f^p_{i0} represents the origin route point of the trajectory and $f^p_{im_i}$ represents the destination route point. To learn the regular flow transfer patterns in flight planning and use them as the prior knowledge, the planned flow of different subregions in all time slices is counted, and the planned flow of k-th subregion at the t_j-th time slice can be calculated as

$$X^{\text{plan}}_{t_j,k} = \sum_{i=1}^{N_f} \sum_{j=0}^{m_i'} y,$$
$$y = \begin{cases} 1, & l_{f^p_{ij}} \in k, T_{f^p_{ij}} \in t_j, \\ 0, & \text{others}, \end{cases} \tag{2}$$

where f^p_{ij} represents j-th plan trajectory point of i-th flight.

Definition 3 (Meteorological Radar Echo). The flow evolution pattern has a strong correlation with external factors such as weather conditions. We devote the weather factor by the meteorological radar echo data, and the preprocessed radar echo data at a certain time step are denoted as a tensor $M_t \in R^{N \times L_w}$, where N is the number of subregion and L_w is weather feature length.

Problem 4 (Air Traffic Flow Prediction). Here, we define the problem of air traffic flow prediction. Given the historical observations of air traffic flow $X^{\text{actual}}_I \in R^{P \times N}$, $I = \{t - P + 1, \ldots t - 1, t\}$, our goal is to predict the air traffic flow in the future time step $X^{\text{actual}}_J \in R^{Q \times N}$, $J = \{t + 1, \ldots, t + Q\}$, where N is the number of regions and P, Q are numbers of historical time intervals and future time intervals, respectively. In this paper, to predict more accurately, the flight plan information $X^{\text{plan}}_{(P+Q),N} \in R^{(P+Q) \times N}$ and meteorological radar echo data $M_{(P+Q)} \in R^{(P+Q) \times N \times L_w}$ in the corresponding time interval are also used.

3.2. Graph Convolutional Networks (GCNs). Spectral graph convolution extends the convolution operation from grid-based data to graph structure data, in which the graph can be represented by its corresponding Laplacian matrix $L \in R^{N \times N}$ [80]. By analyzing the Laplacian matrix $L \in R^{N \times N}$ and its eigenvalues, the properties of the graph structure can be obtained:

$$L = I_n - D^{-(1/2)} W_t D^{-(1/2)}$$
$$= U \Lambda U^T \in R^{n \times n}, \tag{3}$$

where $L \in R^{N\times N}$ is the Laplacian matrix that can represent the graph W_t, I_n is an identity matrix, D is diagonal degree matrix with $D_{ii} = \sum_j W_{ij}$, U is the matrix of the eigenvectors of the normalized graph Laplacian matrix, and Λ is the diagonal matrix of the eigenvalues of L.

Based on the above analysis, the spectral convolution of the air traffic flow $x\epsilon R^{|V|}$ and the kernel Θ on graph W_t can be defined as

$$\begin{aligned}\Theta *_{G(W)} x &= \Theta(L)x \\ &= \Theta\left(U\Lambda U^T\right)x \\ &= U\Theta(\Lambda)U^T x,\end{aligned} \tag{4}$$

where $x\epsilon R^{|V|}$ is signal defined on j-th time slice graph and W_t is the air traffic graph at time j.

However, the computation of the above convolution operation is expensive; in our method, Chebyshev polynomial approximation is adopted to reduce the computation cost of equation (4):

$$\Theta *_{G(W)} x = \Theta(L)x \approx \sum_{k=0}^{K_s-1} \theta_k T_k(\tilde{L})x, \tag{5}$$

where $\theta_k \in R^k$ is a vector of polynomial coefficients. K_s is the kernel size of graph convolution, which determines the maximum radius of the convolution from central nodes. $\tilde{L} = 2L/\lambda_{\max} - I_n$, $\lambda_{\max}$ is the maximum eigenvalue of the Laplacian matrix. $T_k(\tilde{L}) \in R^{N\times N}$ is the Chebyshev polynomial of order k.

3.3. Temporal Gate Convolution. The temporal gate convolutional layer contains a one-dimensional causal convolution with a width of K_t kernel, followed by a gated linear unit [9]. Suppose $y\epsilon R^{T\times c_{in}}$ is the signal defined on i-th node; by temporal convolution of the air traffic flow $y\epsilon R^{T\times c_{in}}$ and the kernel $\Gamma \in R^{K_t\times c_{in}\times 2c_{out}}$, we can obtain output $Y \in R^{(T-K_t+1)\times 2c_{out}}$. Split Y in half with the same size of channels, and $Y_1, Y_2 \in R^{(T-K_t+1)\times c_{out}}$ can be obtained. The temporal convolutional layer can be defined as

$$\Gamma *_T y = Y_1 \odot \sigma(Y_2), \tag{6}$$

where $\odot$ denotes the elementwise Hadamard product and $\sigma(\bullet)$ represents the sigmoid function.

4. Methodology

To fully utilize the unique characteristics of air traffic operation patterns and address the complex air traffic flow prediction problem, we propose a novel Spatiotemporal Knowledge Distillation Network (ST-KDN), the overall architecture of which is illustrated in Figure 2. Recognizing the inherent predictive capabilities embedded within flight plans, encompassing details on future traffic flow between nodes and implicitly encoding dependencies of upstream and downstream flows, we design a teacher network integrating the prior knowledge from flight plans. This network not only learns traffic evolution patterns from historical flow data but also derives anticipatory guidance from future flight plans, facilitating more precise predictions of future traffic patterns. Moreover, considering the significant impact of external factors like thunderstorms on air traffic operations, we propose a student network structured around a "parallel-fusion" design. This architecture segregates the modeling of spatial-temporal dependencies and weather impacts before merging them. Finally, by distilling insights from the teacher network and integrating meteorological features, the student network adeptly captures the intricate spatial-temporal dependency relationships within air traffic, while explicitly simulating the effects of weather on air traffic flow.

4.1. Teacher Network for Prior Knowledge Learning from Flight Plan. In contrast to road traffic, air traffic must adhere to predetermined routes and comply with air traffic controllers' directives to ensure safety, rendering the transfer pattern of air traffic flow intricate and constrained. Flight plans constitute a crucial component of air traffic flow evolution, as they enable the anticipation of flight intentions in advance and furnish insights into how traffic transitions between nodes, thus supplying valuable prior knowledge for future air traffic flow at each node. Consequently, we propose harnessing a teacher network to glean the evolution pattern of rules as significant prior knowledge. Unlike conventional methods that solely rely on historical data for learning traffic patterns, our proposed teacher network integrates insights from flight plans, thereby offering valuable guidance for air traffic prediction.

Given the planned flow of different subregions in historical P time steps $X_{I'}^{\text{plan}} \in R^{P\times N}$, $I' = \{t-P+1, \ldots t-1, t\}$, and the planned flow in future Q time steps, i.e., $L^T = X_{J'}^{\text{plan}} \in R^{Q\times N}$, $J' = \{t+1, \ldots, t+Q\}$. A teacher network consisting of a spatiotemporal feature extraction module is designed to model the regular evolution pattern of planned air traffic flow, as defined in the following equation:

$$O^T = f_{\text{Teacher}}\left(X_{I'}^{\text{plan}}\right), \tag{7}$$

where $O^T \in R^{Q\times N}$ represents the output of the teacher network.

By the teacher network, the regular evolution pattern contained in the flight plan is learned. Specifically, the teacher network consists of two spatiotemporal convolution blocks and a prediction layer. Each spatiotemporal convolution block is composed of two temporal gated convolution layers and a spatial graph convolution layer, and the details of graph convolution and temporal gated convolution are described in the section of Preliminaries. The prediction layer is composed of two temporal gate convolution layers and a fully connection layer. In spatial convolution, the graph convolution operator $*_{G(W)}$ defined on $x\epsilon R^N$ can be extended to multidimensional tensors. For a signal with C_i channels $\in R^{N\times C_i}$, graph convolution in (5) can be generalized as

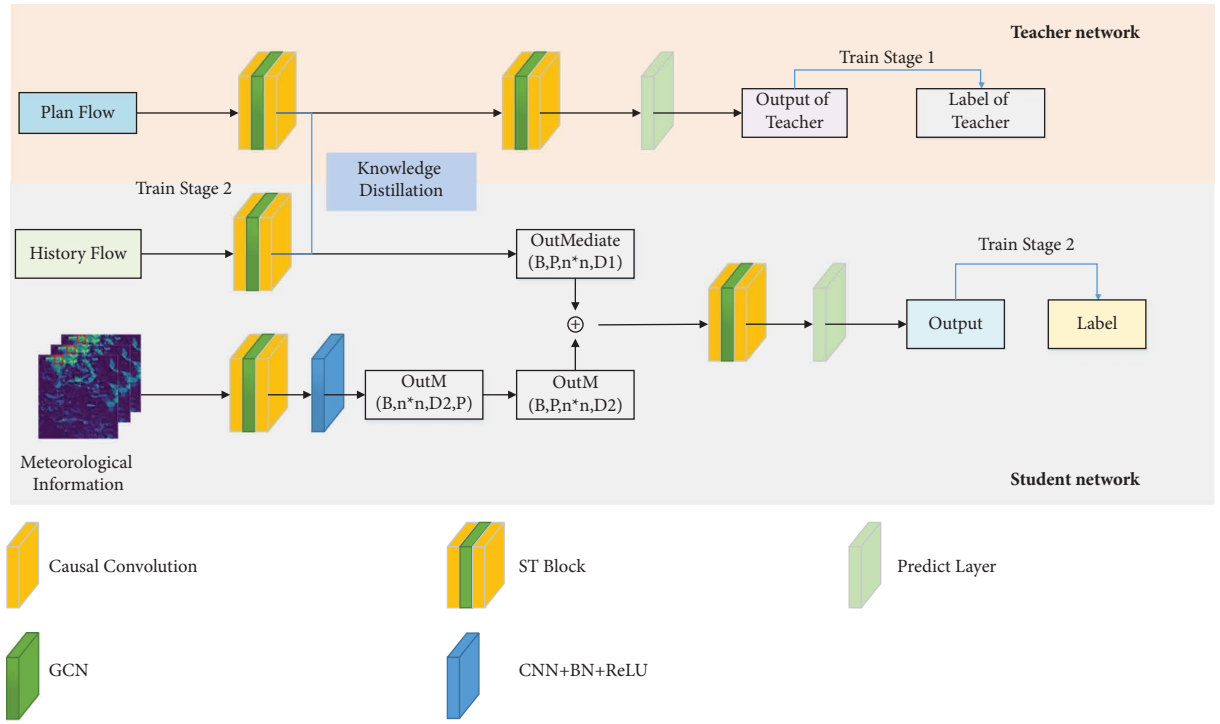

FIGURE 2: The framework of the proposed ST-KDN network. GCN: graph convolution network; ST block: spatiotemporal block; CNN: convolutional neural network; BN: batch normalization.

$$y_j = \sum_{i=1}^{C_i} \theta_{i,j}(L)x_i \in R^N, \quad 1 \le j \le C_o, \tag{8}$$

where $C_i \times C_o$ vectors of Chebyshev coefficients $\theta_{i,j} \in R^K$ and C_i, C_o represent the size of input and output of the feature maps, respectively. The graph convolution for 2-D variables is denoted as "$\theta *_{G(W)} X$" with $\in R^{K \times C_i \times C_o}$. Specifically, the input of traffic prediction is composed of P frame of graphs. Each frame can be regarded as a matrix whose column i is the C_i-dimensional value of each frame at the i th node in graph, as $X \in R^{N \times C_i}$ (in this case, $C_i = 1$). For each time step t of P, the equal graph convolution operation with the same kernel θ is imposed on $X_t \in R^{N \times C_i}$ in parallel. For temporal dimensional features, temporal gate convolution is used. Given i-th node signal $y \in R^{T \times c_{in}}$ and the kernel $\Gamma \in R^{K_t \times c_{in} \times 2c_{out}}$,

$$\Gamma *_T y = f_c(y) \odot \sigma(f_c(y)), \tag{9}$$

where f_c represents one-dimensional causal convolution with a width of K_t kernel, $\odot$ denotes the elementwise Hadamard product, and $\sigma(\bullet)$ is the sigmoid function.

The prior knowledge from the flight plan is embedded in the parameters of the teacher network. To distill knowledge of teacher network to guide the training of the student network, after the first spatiotemporal convolution block, an intermediate output $O_M^T \in R^{P \times N \times D_1}$ of the teacher network is generated.

4.2. Student Network for Learning Nonrecurrent Flow Patterns with Weather Factor. The teacher network has learned the flow evolution pattern implied in the flight plan data, which helps to infer the regular flow evolution pattern without sudden factor disturbance. However, in practice, actual air traffic flow is significantly influenced by meteorological conditions. Given the considerable uncertainty linked with external factors, it is a very challenging problem to explicitly model the impact of weather on air traffic flow and to predict more accurately the nonrecurrent flow patterns.

To address this challenge, a student network based on the "parallel-fusion" structure is proposed. The student network is firstly divided into two parts, which can learn the regular flow evolution pattern and weather change characteristics, respectively. Subsequently, a feature fusion module is designed to integrate the regular flow feature and weather feature, which can explicitly model complex nonrecurrent spatiotemporal dependencies. Based on the above observation, our student model consists of a regular pattern learning module, a weather feature extraction module, and a feature fusion module, as defined in the following equation:

$$Y = f_{ST}\left(f_{ST}\left(X^{actual}\right) \| F_{CNN}\left(f_{ST}(M)\right)\right), \tag{10}$$

where $X^{actual} \in R^{P \times N}$ represents real flow matrix, $M \in R^{T \times N \times L_w}$ represents meteorological radar echo matrix, L_w is weather feature length, f_{ST} represents spatiotemporal convolution block, F_{CNN} is convolutional neural network layer, and $Y \in R^{Q \times N}$ is output flow. It is noteworthy that $T = P + Q$, which means the input of the regular pattern learning module is the flow matrix in historical P time steps, while the input of the weather feature extraction module is the meteorological radar echo matrix consisting of the historical P time steps and the future Q time steps. This is because the flow at the next Q

time steps is highly dependent on both historical and future meteorology conditions.

Considering that air traffic flow distribution of any region has significant dependencies with its neighbors and the flight flow of a certain region is related to its previous observations, the proposed regular pattern learning module consists of spatiotemporal convolution block, i.e., two temporal gated convolution layers and a spatial graph convolution layer. The details of temporal gated convolution layers and spatial graph convolution layer are described in the section of Preliminaries. In spatial graph convolution layer, the adjacency matrix of the air traffic graph is computed based on the distances among subregions. The weighted adjacency matrix $W_t \in R^{n\times n}$ can be formed as

$$W_{t,ij} = \begin{cases} \exp\left(-\frac{d_{ij}^2}{\delta^2}\right), & i \neq j,\ \exp\left(-\frac{d_{ij}^2}{\delta^2}\right) \geq \varepsilon, \\ 0, & \text{others}, \end{cases} \tag{11}$$

where $W_{t,ij} \in R^1$, d_{ij}^2 represents distance between subregions and δ^2 and ε are thresholds to control the distribution and sparsity of matrix $W_t \in R^{n\times n}$, respectively. It is worth noting that the knowledge of regular pattern learning module comes from the teacher network. To capture the spatiotemporal dependencies of meteorological changes, the meteorological feature extraction module is designed, which consists of ST block and convolutional layer. Finally, the feature fusion module explicitly models the nonrecurrent flow patterns affected by weather and makes the final prediction by integrating the regular flow feature and weather feature.

4.3. Knowledge Distilling for Air Traffic Flow Prediction. To better predict real flow affected by weather while learning teacher network knowledge, a feature-based distillation approach is adopted. Compared with the response-based distillation using output layer features, feature-based distillation better utilizes intermediate layer features, thus enhancing the training effectiveness of the student model.

Given the intermediate output of the teacher network $O_M^T \in R^{P\times N\times D_1}$ and the output of the regular pattern learning module in the student network $O_M^S \in R^{P\times N\times D_1}$, we propose to distill the knowledge by approximating $O_M^T \in R^{P\times N\times D_1}$ and $O_M^S \in R^{P\times N\times D_1}$. However, direct fitting the intermediate characteristics may cause the student network to overfit the teacher network, thereby losing other useful information. Therefore, we propose to map $O_M^T \in R^{P\times N\times D_1}$ to the latent space and then distill in the latent space. Thus, the distillation loss between teacher and student can be derived as follows:

$$\mathscr{L}_{\text{distill}} = \frac{1}{N}\sum_{i=1}^{N}\left(F_{\text{CNN}}\left(O_{\text{Mi}}^T\right) - O_{\text{Mi}}^S\right)^2, \tag{12}$$

where $O_{Mi}^T \in R^{P\times 1\times D_1}$, $O_{Mi}^S \in R^{P\times 1\times D_1}$ represent the output value of $O_M^T \in R^{P\times N\times D_1}$, $O_M^S \in R^{P\times N\times D_1}$, respectively.

During training, the whole process is divided into two stages. In the first stage, the teacher model is first trained, and teacher loss $\mathscr{L}_T$ based on mean square error is used:

$$\mathscr{L}_T = \frac{1}{N}\sum_{i=1}^{N}\left(O_i^T - L_i^T\right)^2, \tag{13}$$

where $O_i^T \in R^{1\times T}$ represents prediction values of teacher network on i-th region and $L_i^T \in R^{1\times T}$ represents label of teacher network on i-th region.

In the second stage, the teacher model is used to guide the training of the student network. The loss function for optimizing the student network is given in two parts:

$$\mathscr{L}_{\text{total}} = \alpha \cdot \mathscr{L}_{\text{distill}} + \mathscr{L}_S, \tag{14}$$

in which

$$\mathscr{L}_S = \frac{1}{N}\sum_{i=1}^{N}\left(O_i^S - L_i^S\right)^2, \tag{15}$$

where α is weight adjustment factor and $O_i^S \in R^{1\times T}$, $L_i^S \in R^{1\times T}$ represent prediction value and label value of the student network on the i-th region, respectively.

5. Experiments

In this section, we firstly outline the experimental settings in Section 5.1, which encompasses the datasets, evaluation metrics, and setting of hyperparameters. These details provide necessary background and criteria for understanding the conditions under which our experiments are conducted. Subsequently, in Section 5.2, we introduce the baseline models as comparative benchmarks against our proposed method. Section 5.3 presents experiments for model comparison and variant comparison, focusing on quantitative objective measurements of model performance. Model comparison entails comparisons of prediction errors and time cost, verifying the superiority of the proposed method over other representative methods, while variant comparison aims to validate the effectiveness of each key module within the proposed method. Finally, in Section 5.4, a case study is conducted, emphasizing visually demonstrable results to showcase the performance of ST-KDN in practical scenarios. Furthermore, the discussion in Section 5.5 outlines future research directions. These series of experimental designs are aimed at comprehensively evaluating the performance of our proposed method and clearly demonstrating its advantages and applicability.

5.1. Experiment Settings

5.1.1. Datasets. The original data are provided by the Aviation Data Communication Corporation (ADCC), China. They mainly include trajectory data and meteorological radar echo data, covering the period from May 1, 2021, to July 1, 2021. The trajectory data are composed of flight mission ID, planned/real flight departure information, planned/real route point names, latitude and longitude, flight level, speed, and the corresponding time.

The meteorological radar echo data monitor the weather that affects the flight such as precipitation and strong convective weather. The national airspace contains numerous points, and each kind of data represents the weather conditions where the point is located.

To evaluate the effect of the proposed model, the data are further divided into three parts: 70% for training, 10% for validation, and 20% for testing.

5.1.2. Evaluation Metrics. To demonstrate the effectiveness of the proposed ST-KDN, three widely used metrics are applied, i.e., Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), which are defined as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \widehat{Y}_i|, \tag{16}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}, \tag{17}$$

$$\mathrm{MAPE} = \sum_{i=1}^{n} \left|\frac{Y_i - \widehat{Y}_i}{\widehat{Y}_i}\right| \times \frac{100\%}{n}, \tag{18}$$

where n is the number of testing samples and Y_i and $\widehat{Y}_i$ denote the predicted value and the ground truth of air traffic flow, respectively.

5.1.3. Hyperparameters. In our model, historical time step P is set to 18, which represents 3 hours, and future time step Q is 6 for one hour. We use Adam (Adaptive Moment Estimation) optimization algorithm, the initial learning rate is 0.001, and the batch size is 64. In spatiotemporal graph convolution, the kernel size of spatial convolution K_s is set to 3, the temporal convolution kernel K_t is 3, and δ^2 and ε are assigned to 10 and 0.5. In the weather feature extraction module, L_w is set to 200.

5.2. Baselines. We compare our model with the following six baselines:

(i) SVR [32]: support vector regression is a machine learning model that does not consider spatial correlation.

(ii) ASTGCN [11]: it is an attention-based spatial-temporal graph convolutional network model to predict traffic flow.

(iii) AGCRN [12]: it is a traffic flow prediction method based on the adaptive adjacency matrix graph convolution.

(iv) GMAN [40]: it is a graph multiattention network for traffic prediction.

(v) STSSL [14]: it is a spatiotemporal self-supervised traffic prediction network that considers both spatial and temporal heterogeneities, in which the adaptive heterogeneity-aware data augmentation method is devised on spatiotemporal graphs to mitigate noise disturbances.

(vi) TESTAM [21]: it is a time-enhanced spatiotemporal attention network, primarily integrating temporal characteristics of traffic networks for traffic prediction.

5.3. Experimental Results. To thoroughly ascertain the advantages of the proposed ST-KDN method, this section provides a detailed analysis of the quantitative error results, objectively measuring model performance through specific evaluation metrics. Our study comprises two distinct parts. Firstly, to validate the superiority of the proposed method over other representative methods, we compare the proposed method with other benchmark methods in terms of prediction errors across different time intervals and time cost. Secondly, we conduct variant comparisons of the proposed ST-KDN method to validate the effectiveness of the ST-KDN's key modules. These two types of comparison experiments comprehensively validate the effectiveness of the proposed ST-KDN method from both the overall superiority of the model and the effectiveness of the modules within the proposed model.

5.3.1. Model Comparisons. This section entails comparisons of prediction errors and time cost. Initially, we present comparisons of prediction errors across different time intervals between the proposed method and other benchmark methods to demonstrate its capability in predicting air traffic flow. Subsequently, we compare the time complexity and actual inference time of different models to show the operational efficiency of different methods.

Table 1 shows prediction performance of seven different methods on the real dataset in the next 10 minutes ($Q = 1$), 20 minutes ($Q = 2$), 30 minutes ($Q = 3$), 40 minutes ($Q = 4$), 50 minutes ($Q = 5$), and 60 minutes ($Q = 6$). Overall, as the prediction interval increases, the corresponding prediction difficulty becomes greater, and hence the prediction error is also increasing. As shown in Table 1, we can observe the following results. (1) Deep learning methods are superior to traditional machine learning methods, such as SVR, which proves the powerful ability of neural networks in modeling nonlinear and complex air traffic data. (2) Our model achieves the state-of-the-art prediction performance in most time intervals. This may be because our model integrates prior knowledge of future flow evolution pattern in flight plan and considers weather impact, revealing the effectiveness of modeling weather feature and regular prior knowledge. The proposed "Teacher-Student" framework aids us in obtaining richer priors of future air traffic flow and capturing the non-recurrent dynamics conditions. (3) The proposed method demonstrates particularly pronounced advantages within the long-term prediction horizon (e.g., 40 min, 50 min, and 60 min), aiding in mitigating the error propagation issue across prediction time steps. Long-term traffic prediction

Table 1: The experimental results of the proposed ST-KDN and other six comparison methods in the next 10 minutes, 20 minutes, 30 minutes, 40 minutes, 50 minutes, and 60 minutes.

Metric	Method	10 min	20 min	30 min	40 min	50 min	60 min
MAE	SVR	1.898696	2.244607	2.484253	2.767143	3.052133	3.309817
	AGCRN	1.794790	1.869949	1.940945	2.007488	2.073016	2.138081
	ASTGCN	1.795160	1.890413	1.962362	2.056936	2.060439	2.158814
	GMAN	1.805774	1.798546	1.801262	1.819169	1.848141	1.886582
	STSSL	**1.638903**	1.804966	1.980573	2.128636	2.288099	2.457236
	TESTAM	1.841422	1.857893	1.883490	1.924575	1.977991	1.989194
	ST-KDN	1.763453	**1.770182**	**1.771767**	**1.778287**	**1.779059**	**1.781851**
RMSE	SVR	2.565081	2.983550	3.265532	3.604007	3.952299	4.265697
	AGCRN	2.519696	2.634144	2.735415	2.832605	2.927571	3.023671
	ASTGCN	2.495914	2.630218	2.737751	2.866813	2.889380	3.003969
	GMAN	2.518119	**2.509231**	2.513937	2.539927	2.584160	2.645740
	STSSL	**2.375745**	2.645554	2.916242	3.149616	3.423993	3.716680
	TESTAM	2.563212	2.595453	2.638512	2.712400	2.807796	2.912532
	ST-KDN	2.500712	2.509924	**2.506504**	**2.512559**	**2.511632**	**2.514018**
MAPE	SVR	0.387429	0.459490	0.513001	0.576568	0.640230	0.697416
	AGCRN	0.368877	0.385325	0.404844	0.423151	0.441527	0.460654
	ASTGCN	0.370427	0.385325	0.404844	0.423151	0.441527	0.460654
	GMAN	0.384672	0.379004	0.376588	0.377863	0.382285	0.389442
	STSSL	**0.318166**	**0.333431**	0.377861	0.381922	0.399758	0.408216
	TESTAM	0.379114	0.380991	0.388400	0.391331	0.397276	0.401991
	ST-KDN	0.372582	0.373806	**0.374326**	**0.376572**	**0.375481**	**0.376291**

The optimal results for each index within each prediction interval are indicated by bold values. Table shows prediction performance of seven different methods on the real data set in the next 10 minutes ($Q = 1$), 20 minutes ($Q = 2$), 30 minutes ($Q = 3$), 40 minutes ($Q = 4$), 50 minutes ($Q = 5$), and 60 minutes ($Q = 6$).

presents significant challenges due to the complexity of the transportation system and the myriad influencing factors stemming from the continually changing natural environment. Compared to other methods that primarily focus on spatiotemporal modeling, our method benefits from guidance provided by prior knowledge from flight plans, enabling the capture of valuable priors in challenging long-term predictions. Consequently, our method exhibits more satisfactory performance in the long-term horizon. We argue that the long-term traffic prediction is more beneficial to practical applications, e.g., it allows air traffic controller to have more time to take actions to optimize the air traffic flow according to the prediction.

Furthermore, it is observed that on certain metrics, such as the 10-minute MAE, STSSL achieves lower prediction errors. However, as the prediction horizon extends, the error of the STSSL method gradually increases, whereas the proposed method maintains lower errors in long-term prediction. This is attributed to our model employing synchronous multistep prediction, where the optimization process considers results across multiple time steps. This allows our method to focus more on the overall prediction performance across all time steps, rather than being limited to individual prediction steps. In contrast, the STSSL method focuses on individual short-term predictions, i.e., single-step prediction, followed by iteratively using predicted values as known values to achieve multistep prediction. This procedure introduces error propagation, leading to accumulated errors in multistep prediction. For the 20-minute RMSE metric, the proposed method achieves the second-lowest with a slight gap compared to GMAN, which could be attributed to the higher sensitivity of RMSE to outliers. During the 20-minute prediction, the occurrence of an outlier in the prediction error of the proposed method results in a slightly inferior performance compared to GMAN when computing the RMSE metric. However, our proposed method outperforms GMAN in all other metrics at all other time intervals. This suggests that the proposed knowledge distillation framework is still more advantageous for multistep prediction of air traffic flow than GMAN. We contend that long-term prediction holds greater practical significance as it provides valuable insights for decision-making processes such as air traffic management and resource allocation. Thus, the advantage of our approach in long-term prediction underscores its heightened practical utility.

In addition, we compare the time complexity and actual running time of different models in terms of floating-point operations (FLOPS) and inference time. Typically, FLOPS is a factor employed by many researchers to quantify the time complexity of deep learning algorithms. Smaller FLOPS value indicates lower computational complexity required. However, FLOPS may not reflect the "actual" execution speed of methods as they do not account for algorithm parallelism. As another evaluation metric, inference time can validate the execution speed of model. Here, we record the inference time for each batch of the test set (batch size is 64). All experiments are run on an Intel(R) Xeon(R) Gold 5218 CPU computer with a frequency of 2.30 GHz, using the NVIDIA GeForce RTX 3090 GPU. The programming language used is Python 3.8 and the deep learning framework utilized is PyTorch 1.11.0. Table 2 illustrates the FLOPS and inference time of the proposed ST-KDN and six other comparison methods. It is evident that the proposed ST-KDN achieves the minimum FLOPS compared to other

Table 2: The FLOPS and inference time of the proposed ST-KDN and other six comparison methods.

	FLOPS/GFLOPS	Inference time/millisecond, batch_size = 64
SVR	—	17.8853225
AGCRN	3.15	11.6408624
ASTGCN	2.68	12.6530639
GMAN	25.67	16.5718712
STSSL	3.26	14.6473204
TESTAM	4.67	15.9177380
ST-KDN	2.22	12.4769866

methods. About inference time, the lowest value is the AGCRN method with 11.6408624, but our method achieves the second-lowest inference time with a little gap. This demonstrates the effectiveness of our approach in terms of operational efficiency. As a supplementary clarification, we solely present the FLOPS values associated with diverse deep learning methodologies. This choice is due to significant computational strategy disparities between deep learning approaches and SVR. While deep learning methods typically involve substantial floating-point operations, SVR does not. Consequently, comparing FLOPS between SVR and deep learning methodologies could be misleading.

5.3.2. Variant Comparison. To verify the effectiveness of the various components within the proposed ST-KDN framework, we conduct ablation experiments on a real dataset. These experiments aim to systematically evaluate the contributions of individual components, namely, weather feature modeling and teacher network guidance, to the overall performance of the model. For convenience, we call the model that removes the teacher network guidance as ST-KDN-NT and the model that removes the teacher network guidance and weather feature extraction modules simultaneously as ST-KDN-NTW. By comparing the results of the ablation experiments, the effectiveness of weather feature modeling and teacher network guidance has been proved.

(1) Effect of Weather Feature Modeling. Figure 3(a) shows the comparative experimental results of ST-KDN-NT and ST-KDN-NTW. It can be seen from Figure 3(a) that the ST-KDN-NT model can obtain a lower prediction error than ST-KDN-NTW in all prediction intervals, which shows that the weather extraction module helps to capture more complex spatiotemporal dependencies.

(2) Effect of Teacher Network Guidance. To investigate the effectiveness of the proposed ST-KDN, we compare the prediction results of the proposed ST-KDN and the ST-KDN-NT, as shown in Figure 3(b). We observe that the proposed ST-KDN can achieve better prediction performance. This may be because the regular flow transfer pattern from the flight plan is learned, which can provide effective prior knowledge in air traffic flow. Through the guidance of the teacher network and the extraction of meteorological features, the student network can explicitly model the impact of weather on air traffic.

5.4. Case Study. To provide visual and intuitive insights about the effectiveness of the proposed method, as well as to further validate its advantages in learning nonrecurrent air traffic flow patterns influenced by weather, we present a series of visualization examples. These illustrations aid readers in gaining a comprehensive understanding of our research outcomes and provide them with intuitive visual impressions. Figure 4 compares the prediction errors of two regions where thunderstorms exist for 60 consecutive minutes. Five air traffic prediction methods based on deep learning, i.e., GMAN, ASTGCN, AGCRN, STSSL, and TESTAM, and two variants, i.e., ST-KDN-NT and ST-KDN-NTW, are used. In Figure 4, the first row shows the meteorological radar echo maps for 60 minutes with 20 min interval, where the darker yellow point represents the regions with more severe thunderstorm. We select three representative regions and frame them in red. For the red box region, the prediction error of different methods with a time interval of 20 minutes is displayed in the second row in Figure 4. From Figure 4, we can see the following. (1) Compared to ST-KDN-NT and ST-KDN-NTW, the proposed ST-KDN method achieves lower prediction errors, indicating the significant role of "weather feature modeling" and "teacher network guidance" in effectively modeling the impact of weather on air traffic flow. (2) From Figures 4(a) and 4(b), it is evident that the proposed ST-KDN outperforms other traffic prediction methods during severe thunderstorms, highlighting its significant advantage in modeling nonrecurrent spatiotemporal dependencies.

5.5. Discussion. In this section, we discuss the limitations of the proposed method by presenting failure visual examples, as shown in Figure 5. It can be observed that within the initial 20-minute prediction interval, the STSSL method achieves the lowest prediction error, followed by the proposed ST-KDN method. This can be attributed to our model's adoption of synchronous multistep prediction, which simultaneously considers the prediction errors of multiple consecutive time steps, whereas the STSSL method focuses solely on single-step prediction performance. However, despite this, within the 40-minute and 60-minute prediction intervals, the proposed ST-KDN consistently achieves the lowest prediction errors. In the future, we aim to explore methodologies that effectively reconcile the prediction efficacy between long-term and short-term perspectives.

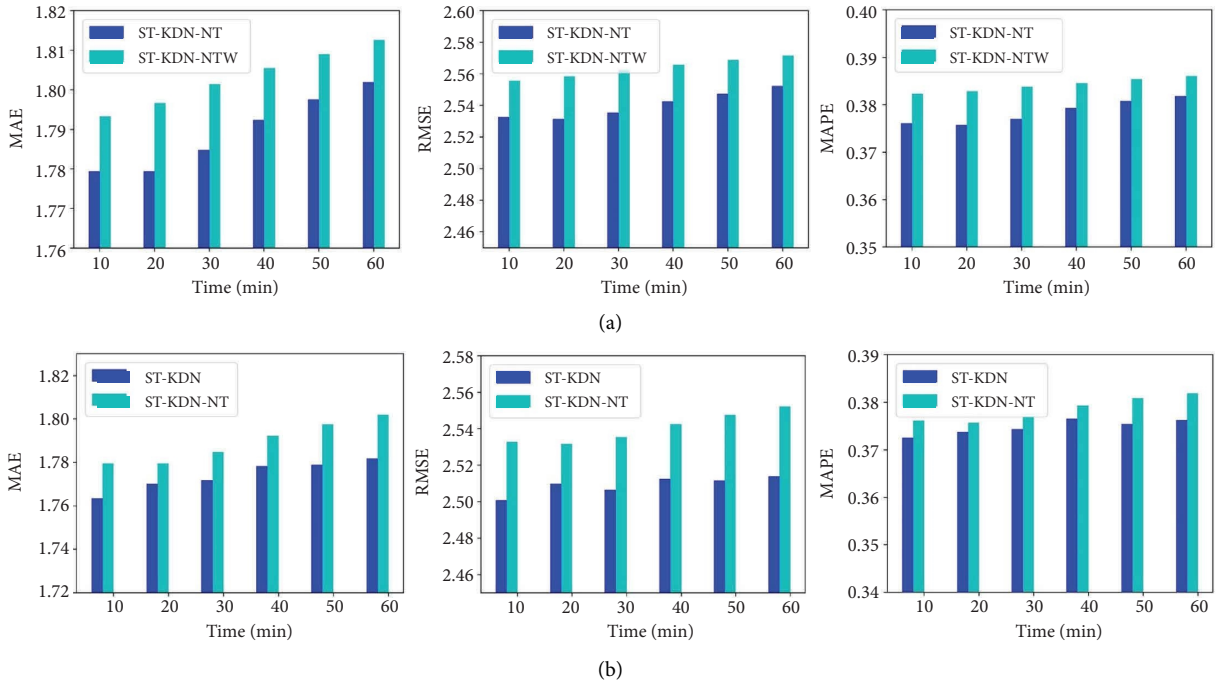

Figure 3: (a) The comparative experimental results of ST-KDN-NT and ST-KDN-NTW to investigate effect of weather feature modeling. (b) The comparative experimental results of ST-KDN-NT and the proposed ST-KDN to investigate effect of teacher network guidance.

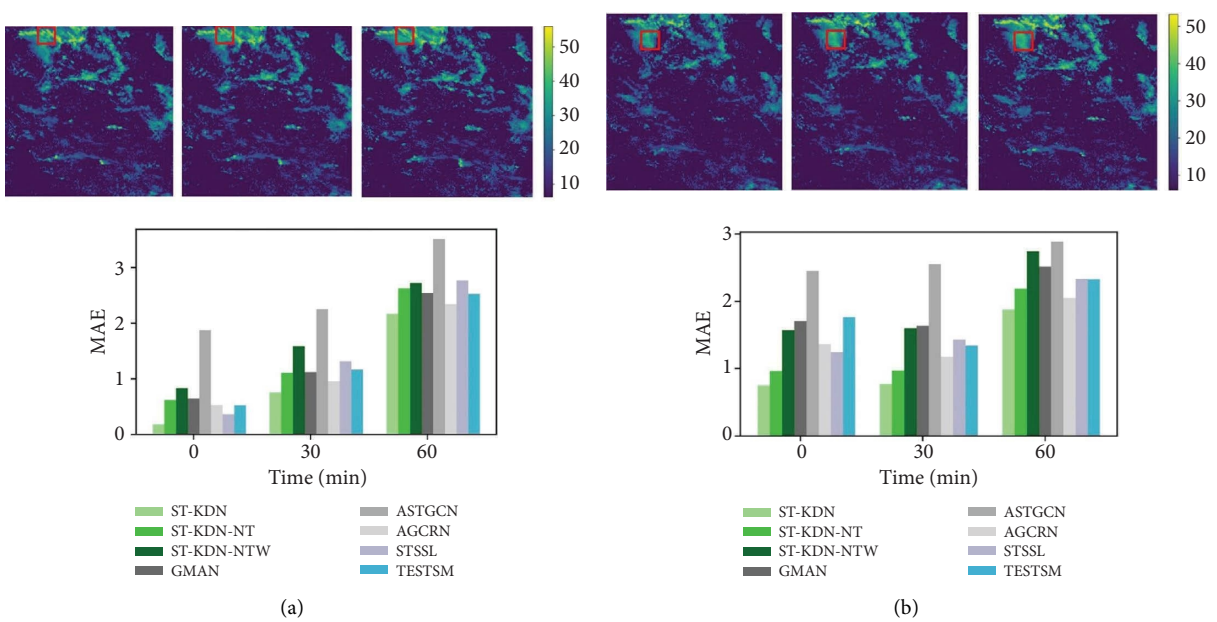

Figure 4: The comparison of prediction errors of different methods for 60 consecutive minutes in two thunderstorm regions. (a) Region I. (b) Region II.

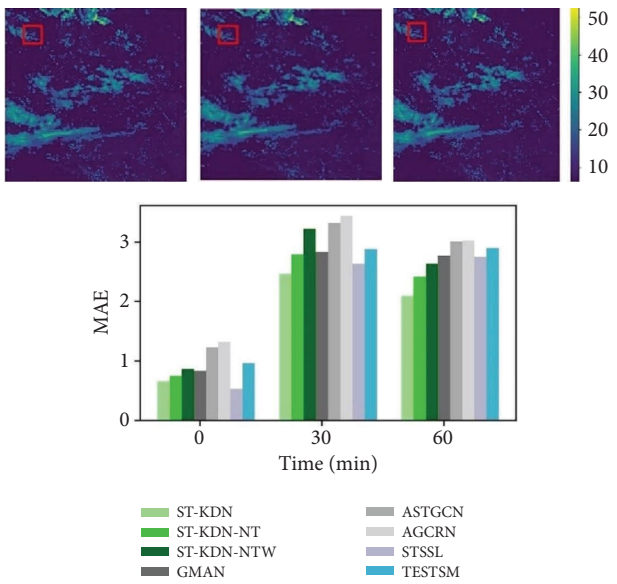

Figure 5: The comparison of prediction errors of different methods for 60 consecutive minutes in a region with less severe thunderstorm.

6. Conclusion

In this paper, we propose a spatiotemporal knowledge distillation network for air traffic flow prediction. A "teacher-student" distillation model considering flight plan prior is designed to integrate prior knowledge of regular flow evolution pattern into learning network. The teacher network is used to learn regular air traffic pattern from flight plans as a priori. Based on this, a student network based on a "parallel-fusion" structure is proposed, and the student network consists of a regular pattern learning module, a weather feature extraction module, and a feature fusion module. The regular pattern learning module learns the knowledge from teacher network, the weather feature extraction module mines meteorological features that have an impact on the air traffic flow, and the nonrecurrent air flow evolution pattern is modeled by feature fusion module. By knowledge distillation and meteorological feature extraction, our method explicitly models nonrecurrent spatiotemporal dependencies. The experimental results on real-world flight data demonstrate that the proposed method could effectively capture rules of airspace flow variation and achieve a better prediction performance, especially in predicting air traffic flow affected by thunderstorms.

Data Availability

The data used to support the findings of this study are available from the corresponding author on request. The data are not publicly available due to privacy.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Funds of the National Natural Science Foundation of China (grant nos. U2133210 and U2033215). The authors are grateful for this support. Furthermore, the authors extend their gratitude to the Aviation Data Communication Corporation (ADCC) in China for their invaluable support in providing the air traffic flow data.

References

[1] J. D. Watson and F. H. C. Crick, "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737-738, 1953.

[2] W. Wang, K. Cai, W. Du et al., "Analysis of the Chinese railway system as a complex network," *Chaos, Solitons & Fractals*, vol. 130, Article ID 109408, 2020.

[3] O. Idrissi, A. Bikir, and K. Mansouri, "Improving the management of air traffic congestion during the approach phase," *Aeronautical Journal*, vol. 127, no. 1316, pp. 1752–1773, 2023.

[4] J. C. Jones, Z. Ellenbogen, and Y. Glina, "Recommending strategic air traffic management initiatives in convective weather," *Journal of Air Transportation*, vol. 31, no. 2, pp. 45–56, 2023.

[5] X. Dai, M. Hu, W. Tian, and H. Liu, "Modeling congestion propagation in multistage schedule within an airport network," *Journal of Advanced Transportation*, vol. 2018, Article ID 6814348, 11 pages, 2018.

[6] D. Chen, M. Hu, Y. Ma, and J. Yin, "A network-based dynamic air traffic flow model for short-term en route traffic prediction," *Journal of Advanced Transportation*, vol. 50, no. 8, pp. 2174–2192, 2016.

[7] A. Bayen, P. Grieder, and C. Tomlin, "A control theoretic predictive model for sector-based air traffic flow," *AIAA Guidance, Navigation, And Control Conference And Exhibit*, vol. 5011, 2002.

[8] S. Chen, "Short-term air traffic flow prediction based on runtime data of Eurocat-X system," *Information/Communication*, vol. 5, pp. 42–44, 2013.

[9] Y. Lin, J. Zhang, and H. Liu, "Deep learning based short-term air traffic flow prediction considering temporal–spatial correlation," *Aerospace Science and Technology*, vol. 93, Article ID 105113, 2019.

[10] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18*, pp. 3634–3640, Beijing, China, July 2018.

[11] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 922–929, 2019.

[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] J. Jiang, C. Han, W. X. Zhao, and J. Wang, "PDFormer: propagation delay-aware dynamic long-range transformer for traffic flow prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, pp. 4365–4373, 2023.

[14] J. Ji, J. Wang, C. Huang et al., "Spatio-temporal self-supervised learning for traffic flow prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, pp. 4356–4364, 2023.

[15] Y. Xiao, J. J. Liu, J. Xiao, Y. Hu, H. Bu, and S. Wang, "Application of multiscale analysis-based intelligent ensemble modeling on airport traffic forecast," *Transportation Letters*, vol. 7, no. 2, pp. 73–79, 2015.

[16] B. Liao, J. Zhang, C. Wu et al., "Deep sequence learning with auxiliary information for traffic prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 537–546, London, UK, July 2018.

[17] Y. Liang, K. Ouyang, J. Sun et al., "Fine-grained urban flow prediction," in *Proceedings of the Web Conference 2021*, pp. 1833–1845, New York, NY, USA, June 2021.

[18] J. Zhu, Q. Wang, C. Tao, H. Deng, L. Zhao, and H. Li, "AST-GCN: attribute-augmented spatio-temporal graph convolutional network for traffic forecasting," *IEEE Access*, vol. 9, pp. 35973–35983, 2021.

[19] K. Cai, Z. Shen, X. Luo, and Y. Li, "Temporal attention aware dual-graph convolution network for air traffic flow prediction," *Journal of Air Transport Management*, vol. 106, Article ID 102301, 2023.

[20] Y. Zhu and Q. Ni, "A period-extracted multi-featured dynamic graph convolution network for traffic demand prediction," *Applied Intelligence*, vol. 54, no. 1, pp. 722–737, 2024.

[21] H. Lee and S. Ko, "TESTAM: a time-enhanced spatio-temporal attention model with mixture of experts," in *Proceedings of the Twelfth International Conference on Learning Representations*, Vienna, Austria, March 2024.

[22] X. Wang, Y. Shang, and G. Li, "DTM-GCN: a traffic flow prediction model based on dynamic graph convolutional network," *Multimedia Tools and Applications*, pp. 1–17, 2024.

[23] S. E. Said and D. A. Dickey, "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, vol. 71, no. 3, pp. 599–607, 1984.

[24] B. M. Williams, P. K. Durvasula, and D. E. Brown, "Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1644, no. 1, pp. 132–141, 1998.

[25] E. Cadenas, W. Rivera, R. Campos-Amezcua, and C. Heard, "Wind speed prediction using a univariate ARIMA model and a multivariate NARX model," *Energies*, vol. 9, no. 2, p. 109, 2016.

[26] S. Mehrmolaei and M. R. Keyvanpour, "Time series forecasting using improved ARIMA," in *Proceedings of the 2016 Artificial Intelligence and Robotics (IRANOPEN)*, pp. 92–97, IEEE, Qazvin, Iran, April 2016.

[27] S. R. Chandra and H. Al-Deek, "Predictions of freeway traffic speeds and volumes using vector autoregressive models," *Journal of Intelligent Transportation Systems*, vol. 13, no. 2, pp. 53–72, 2009.

[28] 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: Journal of the Transportation Research Board*, vol. 2215, no. 1, pp. 85–92, 2011.

[29] M. Li, P. Tong, M. Li, Z. Jin, J. Huang, and X. S. Hua, "Traffic flow prediction with vehicle trajectories," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 294–302, 2021.

[30] Y. Xiao, Y. Ma, and H. Ding, "Air traffic flow prediction based on k nearest neighbor regression," in *Proceedings of the 2018 13th World Congress on Intelligent Control and Automation (WCICA)*, pp. 1265–1269, Changsha, China, July 2018.

[31] G. A. Davis and N. L. Nihan, "Nonparametric regression and short-term freeway traffic forecasting," *Journal of Transportation Engineering*, vol. 117, no. 2, pp. 178–188, 1991.

[32] H. Zhang, C. Jiang, and L. Yang, "Forecasting traffic congestion status in terminal areas based on support vector machine," *Advances in Mechanical Engineering*, vol. 8, no. 9, Article ID 168781401666738, 2016.

[33] F. Qiu and Y. Li, "Air traffic flow of genetic algorithm to optimize wavelet neural network prediction," in *Proceedings of the 2014 IEEE 5th International Conference on Software Engineering and Service Science*, pp. 1162–1165, Beijing, China, June 2014.

[34] Z. Zhu, B. Peng, C. Xiong, and L. Zhang, "Short-term traffic flow prediction with linear conditional Gaussian Bayesian network," *Journal of Advanced Transportation*, vol. 50, no. 6, pp. 1111–1123, 2016.

[35] G. Lin, A. Lin, and D. Gu, "Using support vector regression and K-nearest neighbors for short-term traffic flow prediction based on maximal information coefficient," *Information Sciences*, vol. 608, pp. 517–531, 2022.

[36] A. Ali, Y. Zhu, and M. Zakarya, "Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction," *Neural Networks*, vol. 145, pp. 233–247, 2022.

[37] Z. Diao, X. Wang, D. Zhang, Y. Liu, K. Xie, and S. He, "Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 890–897, 2019.

[38] L. Liu, J. Chen, H. Wu, J. Zhen, G. Li, and L. Lin, "Physical-virtual collaboration modeling for intra-and inter-station metro ridership prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3377–3391, 2022.

[39] H. Liu, Y. Lin, Z. Chen, D. Guo, J. Zhang, and H. Jing, "Research on the air traffic flow prediction using a deep learning approach," *IEEE Access*, vol. 7, pp. 148019–148030, 2019.

[40] C. Zheng, X. Fan, C. Wang, and J. Qi, "Gman: a graph multi-attention network for traffic prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, pp. 1234–1241, 2020.

[41] C. Ma, G. Dai, and J. Zhou, "Short-term traffic flow prediction for urban road sections based on time series analysis and LSTM_BILSTM method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5615–5624, 2022.

[42] X. Huang, Y. Ye, X. Yang, and L. Xiong, "Multi-view dynamic graph convolution neural network for traffic flow prediction," *Expert Systems with Applications*, vol. 222, Article ID 119779, 2023.

[43] T. M. Rajeh, T. Li, C. Li, M. H. Javed, Z. Luo, and F. Alhaek, "Modeling multi-regional temporal correlation with gated recurrent unit and multiple linear regression for urban traffic flow prediction," *Knowledge-Based Systems*, vol. 262, Article ID 110237, 2023.

[44] S. Wang, J. Cao, and P. S. Yu, "Deep learning for spatio-temporal data mining: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3681–3700, 2022.

[45] J. F. Roddick and B. G. Lees, "Paradigms for spatial and spatio-temporal data mining," *Geographic data mining and knowledge discovery*, pp. 33–50, 2001.

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

[47] Y. Yin, Z. Shen, L. Zhang, and R. Zimmermann, "Spatial-temporal tag mining for automatic geospatial video annotation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 11, no. 2, pp. 1–21, 2015.

[48] Y. Liu, K. Wang, G. Li, and L. Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 5573–5588, 2021.

[49] Y. Liu, K. Wang, L. Liu, H. Lan, and L. Lin, "Tcgl: temporal contrastive graph for self-supervised video representation learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 1978–1993, 2022.

[50] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Deep image-to-video adaptation and fusion networks for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 3168–3182, 2020.

[51] K. Zhang, F. Zhou, L. Wu, N. Xie, and Z. He, "Semantic understanding and prompt engineering for large-scale traffic data imputation," *Information Fusion*, vol. 102, Article ID 102038, 2024.

[52] G. Liang, U. Kintak, X. Ning, P. Tiwari, S. Nowaczyk, and N. Kumar, "Semantics-aware dynamic graph convolutional network for traffic flow forecasting," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 7796–7809, 2023.

[53] Y. Zhu, Y. Zhang, L. Liu et al., "Hybrid-order representation learning for electricity theft detection," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1248–1259, 2023.

[54] J. Lin, Y. Zhu, L. Liu, Y. Liu, G. Li, and L. Lin, "DenseLight: efficient control for large-scale traffic signals with dense feedback," *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.

[55] K. Wang, L. B. Liu, Y. Liu, G. B. Li, F. Zhou, and L. Lin, "Urban regional function guided traffic flow prediction," *Information Sciences*, vol. 634, pp. 308–320, 2023.

[56] J. Gou, L. Sun, B. Yu, S. Wan, and D. Tao, "Hierarchical multi-attention transfer for knowledge distillation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 2, pp. 1–20, 2023.

[57] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Computer Science*, vol. 14, no. 7, pp. 38-39, 2015.

[58] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: fast optimization, network minimization and transfer learning," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, HI, USA, June 2017.

[59] J. Ba and R. Caruana, "Do deep nets really need to be deep?" *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[60] C. Bucilua, R. Caruana, and A. N. Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, Kusatsu, Japan, March 2006.

[61] S. K. Kang, J. Hwang, W. Kweon, and H. Yu, "Topology distillation for recommender system," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 829–839, New York, NY, USA, August 2021.

[62] X. Dai, Z. Jiang, Z. Wu et al., "General instance distillation for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7842–7851, Nashville, TN, USA, June 2021.

[63] P. Passban, Y. Wu, M. Rezagholizadeh, and Q. Liu, "ALP-KD: attention-based layer projection for knowledge distillation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, pp. 13657–13665, 2021.

[64] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2827–2836, San Juan, PR, USA, June 2016.

[65] A. Yusuf, C. Vondrick, and A. Torralba, "Soundnet: learning sound representations from unlabeled video," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[66] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 292–301, Oxford, UK, July 2018.

[67] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2018.

[68] K. Clark, M. T. Luong, U. Khandelwal, C. D. Manning, and Q. V. Le, "Bam! born-again multi-task networks for natural language understanding," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5931–5937, Florence, Italy, July 2019.

[69] S. Flennerhag, P. G. Moreno, N. D. Lawrence, and A. Damianou, "Transferring knowledge across learning processes," in *Proceedings of the International Conference on Learning Representations*, London, UK, March 2018.

[70] F. M. Thoker and J. Gall, "Cross-modal knowledge distillation for action recognition," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, pp. 6–10, Taipei, Taiwan, September 2019.

[71] S. Zhao, G. Wang, S. Zhang et al., "Multi-source distilling domain adaptation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12975–12983, 2020.

[72] Z. Lu, J. Chen, D. Lian, Z. Zhang, Y. Ge, and E. Chen, "Knowledge distillation for high dimensional search index," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[73] C. Yang, Z. An, H. Zhou, F. Zhuang, Y. Xu, and Q. Zhang, "Online knowledge distillation via mutual contrastive learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10212–10227, 2023.

[74] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2604–2613, Long Beach, CA, USA, June 2019.

[75] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: hints for thin deep nets," *Computer Science*, 2015.

[76] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," in *Proceedings of the International Conference on Learning Representations*, Paris, France, July 2017.

[77] Z. Li, P. Xu, X. Chang et al., "When object detection meets knowledge distillation: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10555–10579, 2023.

[78] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, "Learning lightweight lane detection cnns by self attention distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1013–1021, Seoul, Korea, October 2019.

[79] X. Zhou, X. Zheng, X. Cui et al., "Digital twin enhanced federated reinforcement learning with lightweight knowledge distillation in mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 10, pp. 3191–3211, 2023.

[80] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of the International Conference on Learning Representations*, Wuhan, China, July 2016.