Correlated Channeled Spatio Temporal Graph Attention Network Model for Traffic Prediction

Vishal Krishna Singh, Neeraj Jain, Gaurav Tripathi, *Member, IEEE*, and Sharad Sahani

***Abstract*— Accurate real-time traffic prediction in intelligent transportation systems, remains a challenging task because of the frequently changing spatio-temporal dependencies, complex dynamics of the network and varying traffic density. While, existing methods have widely explored the hidden linkages for real-time traffic prediction, multi-faceted interactions between various traffic signals can play an important role in improving the accuracy rate of such systems. This work presents a novel method for analyzing the interrelated interactions between different features of traffic signals through a co-related channeled-spatio-temporal graph attention network. The proposed model is aimed to reflect the intricate channel, temporal and spatial connections between various traffic signal components. The model learns three different embeddings to describe the interactions between traffic signals using graph attention networks. These embeddings are in-line with the signals temporal dynamics, spatial interactions, and channel-based features. Correlation score is used to evaluate the degree of similarity between nodes in various time windows, which ultimately aids in the selection of the most important data. The proposed model is able to outperform various state-of-the-art methods in a series of trials on five real-world datasets. A comprehensive analysis of the results prove that the proposed model not only captures the spatio-temporal correlations in traffic patterns, but is also able to take into account the interactions between traffic signals and their impact on traffic flow changes over time, leading to a 9.7% improved error rate.**

***Index Terms*— Attention, co-relation, graph networks, heterogeneity, traffic dynamics, traffic forecasting.**

I. Introduction

INTELLIGENT transport systems (ITS) are aimed at accurate traffic forecasting and are often driven by time-based or location-based traffic signals. The data, fed to these models, is primarily based on factors such as node-based volume, density, speed, etc. These observations provide insight into real-time traffic c onditions w hich a re r ecorded b y sensors, installed at different spatio-temporal locations in the traffic network. ITS, as such, is interdependent on the reported observations steered by the fact that traffic signal's entirety is connected with other traffic signals, both temporally and geographically, to create a coupled traffic network. To clarify, consider the case of a traffic node in the center of such a network. The traffic condition at this node could be connected to those in it's proximity and also to the distant nodes throughout the network, creating spatial relationships between the nodes. Additionally, there may be temporal relationships between the traffic lights at one node and those at the backward and forward nodes. Evidently, the traffic circumstances are node and time evolving much as spatial and temporal interactions. However, the signals are prone and non-resilient to surge in traffic collisions and other exceptional events that happen on the road. It is pertinent to note that for such traffic recommender systems, the spatio-temporal interactions of traffic components and the conditions on the road network may also be explicit and implicit, hierarchical and diverse, and analogous to the interactions between the end-users [1], [2]. Therefore, it is challenging to exactly forecast how the increase in vehicle flow at one node, at one point of time, may affect the change in vehicle flows at other nodes at subsequent time steps due to these multi-aspect interactions and couplings. Precisely modeling the various spatio-temporal traffic signal couplings across channels and nodes is vital and remains an open research problem.

To address these issues, recent years have seen a steep proliferation in data-driven strategies dominating traditional knowledge-based approaches to traffic forecasting, such as those based on queuing theory [3]. Early time-series analytic approaches like Vector Auto Regression (VAR) [4], utilized the assumption of *stationarity*, which served as the fundamental basis for these methodologies. In contemporary research, deep learning (DL) based methods have been instrumental in generating accurate predictions and eliminating the underlying assumption of *stationarity*. For example, the work presented in [5], [6], [7], and [8], have used Recurrent Neural Network (RNNs) and their variants, such as Gated Recurrent Unit (GRU) [9] and Long Short-Term Memory (LSTM) [10], to exploit their exceptional capability in modeling the temporal dynamics of time series data. These methods aim to automatically learn the intrinsic relationships in the traffic data but fail to consider the additional associations which might be present in the spatio-temporal domain. Interestingly, Graph Multi-attention Networks (GMAN) based methods present a unique solution to this issue. These methods utilize the encoder-decoder architecture for projection of the long-term flow of traffic and are able to achieve promising results

Vishal Krishna Singh is with the School of Computer Science and Electronics Engineering, University of Essex, CO4 3SQ Colchester, U.K. (e-mail: v.k.singh@essex.ac.uk).

Neeraj Jain is with the Department of Computer Science and Engineering and Information Technology, Jaypee Institute of Information Technology, Noida 201309, India (e-mail: neeraj.jain@jiit.ac.in).

Gaurav Tripathi is with GE Vernova, Hyderabad 500081, India (e-mail: gt.iiitl@gmail.com).

Sharad Sahani is with AMD, Hyderabad 500081, India (e-mail: mcs21026@iiitl.ac.in).

in terms of prediction accuracy [11]. Using self-attention in conjunction with convolutional operations, attention-based prediction models exhibit some major advantages over other contemporary methods [11], [12], [13]. The idea is to procreate the networks ability to capture spatio-temporal dependencies and thereby improve the overall predictive performance.

Some of the other important DL based methods for accurate forecasting in smart traffic networks include the ones presented in [14], [15], and [16], where Convolutional Neural Networks (CNNs) are used to stress spatial relationships among traffic data collected from different nodes. The Non-Euclidean nature of real-world traffic data inherently disrupts the spatial structure among different traffic series. CNN based algorithms are capable of manipulating regular grid data, such as $2D$ photos, and therefore attempt to impose a euclidean space on the traffic data. Further, to estimate the spatial connections within traffic road networks, Graph Neural Networks (GNNs) like the Graph Convolution Network (GCN) [17], which is a specialized variant of CNN, is used for the analysis of graph-structured data. The STGCN [18] is one such method used as a means of forecasting traffic patterns and employs graph-based structures in lieu of conventional recurrent and convolutional neural units commonly found in DL frameworks. However, considering that the module's input and output are two sequences of vectors, most of existing methods are faced with the challenge of deploying the self-attention module in multi-dimensional data space due to the separate development of the module either in spatial or in the temporal dimension of the traffic data.

Some of the significant contributions, in GNNs based prediction methods, have considered spatial-temporal relations to minimize the error rate and improve the accuracy. For example, consider the Traffic-GGNN algorithm proposed in [19]. The authors use the bidirectional message passing and the GRU-based real-time spatial-temporal fusion method to improve prediction accuracy. Similarly, the GAT-STC [20] captures both the short as well as the long term traffic patterns using a Temporal Convolutional Network (TCN) with a self-attention mechanism and a distance-based Graph Attention Network (GAT) and spatial-temporal clustering algorithm, respectively. However, the high computational cost and limited performance under strong noise, poses severe constraints on the performance when validated against large scale datasets and diverse real-world scenarios. The authors in [21] propose the STGAGRTN model with the aim to achieve high accuracy. The model extracts dynamic spatial dependencies among the participating nodes of the deployed network and uses the transformer modules to address the issues of real-time updates and computational complexity. However, when deployed to large road networks, the algorithm is unable to match the scaling and fails to achieve the expected computational efficiency. Furthermore, with multi-component structure, the cost of implementation and testing is significantly increased. The STAGCN-EC [22] and GSAA [23] are another important methods of the recent years. The models propose an interesting approach on traffic flow prediction using graph learning, however, the issues with achieving high accuracy with low complexity in large scale datasets, remain an important concern.

Existing methods aim to automatically learn the intrinsic relationships in the traffic data but fail to consider the additional associations which might be present in the spatio-temporal domain such as the channel based characteristics. It is empirical to note that accurately forecasting the impact of an increase in vehicle flow at a single node at a given instance, is challenging. This is due to the complex multi-aspect interactions that impose changes in the flow of the vehicle, at other nodes in subsequent time steps. While some of the recent approaches make an effort to utilize the intrinsic channel characteristics, but it is at the cost of high computational complexity. When deployed to large road networks, the existing algorithms are unable to match the scaling and fail to achieve the expected computational efficiency. Furthermore, with multi-component structure, the cost of implementation and testing is significantly increased. To summarize, while existing methods have widely explored the hidden linkages for real-time traffic prediction, multi-faceted interactions between various traffic signals are important for precise modeling of spatio-temporal traffic signal couplings across channels and nodes.

Therefore, with the aim of accurate real-time traffic prediction in ITS, this work employs a GCN for spatial modeling of a transport system as a spatio-temporal graph. The traffic signal readings, captured by various geo-sensors situated in a traffic road network, is used as the 'traffic data'. The idea is to analyze the interrelated interactions between the different features of traffic signals through the proposed *Co-related Channeled Spatio Temporal Graph Attention Network (CCST-GAN)*. In contrast to the existing methods, the proposed CCST-GAN incorporates a time-gated technique for traffic prediction and involves multi-component fusion. This approach makes training sample preparation more flexible, which is a critical factor in determining the final performance and generalization capability of a supervised learning model. A correlation representation for spatial dynamics is defined through a 'maximum match' score and is termed as *MX score*. Further, the intricate channel, temporal and spatial connections between the various traffic signal components are used to precisely depict the relevant properties across temporal sequences. This information is then used to collect comparable patterns from every sensor and create a dynamic correlation information network, termed as the *MX Score based Graph Convolution Network (MXGCN)*. The proposed CCST-GAN assigns weight to the actual significance of the input graphs of the time-series data for each component, generating a more varied output sequence. Unlike existing methods that rely on intuitive judgments of the importance of the input sequence, the proposed approach utilizes the intrinsic channel characteristics to offer a computationally effectively and accurate solution. The proposed CCST-GAN integrates a robust data augmentation mechanism that leverages the entire temporal dimension of the available data to detect the latent patterns that are highly conducive to predictive modeling. A summary of some of the major methods discussed in the literature is presented in Table I. A careful observation of Table I gives a comprehensive overview of the main features of some of the important methods and also highlights the major advantages of the proposed method over the compared approaches. Our analysis focuses on the robustness of the interconnections among nodes and the tomography of the nodes that comprise the framework of the given graph. In summary, the novel contributions of this work are as follows.

1) This work presents the CCST-GAN method, for accurate real-time traffic prediction in traffic networks. Building on the existing works, the CCST-GAN uses the novel MXGCN module for catching comparable patterns from various sensor nodes.
2) The MXGCN module combines structural information with MIC allowing it to provide the propagate-aggregate mode with precise and deep inter-connections. The advantage is seen in terms of the proposed method creating more accurate features for other components.
3) Contrary to the existing methods, the proposed CCST-GAN uses the light-head architecture and multi-dimensional self-attention scheme to reduce the number of parameters, making the model easier to train. The multi-component structure further allows for parallel processing of different time slices, reducing the overall training time and computational cost.
4) The proposed CCST-GAN uses the MXGCN to construct accurate features by associating the propagate-aggregate mode with the predefined graph structure and density similarity representation. Additionally, the proposed CCST-GAN assigns weight to the actual significance of the input graphs of the time-series data for each component, generating a more varied output sequence and, in turn, stabilizing the contextual features.
5) The proposed CCST-GAN makes use of an efficient data augmentation approach, making it a multi-component fusion-based spatial-temporal graph model. Furthermore, the deep framework based on the proposed MX score based self-attention method, is able to differentiate adaptively and synchronously with the resulting traffic situations in practical forecasting scenarios.
6) A comprehensive analysis of the proposed CCST-GAN along with the state-of-the-art algorithms is presented on a wide range of real-world datasets. The analysis aims to validate the performance in terms of accuracy and efficiency of the proposed method.

The rest of the paper is organized as follows: the Section II presents the definition of major concepts for this work along with the problem statement. It is followed by Section III which presents the details about the proposed methodology. The experiments and results are presented in Section IV followed by the concluding comments in Section V.

II. Definitions and Problem Statement

A. Maximum Match Score

A correlation representation for spatial dynamics is defined through a 'maximum match' score, termed as *MX score*. The MX score acts a means of indicating the interdependence between channeled-spatio-temporal sensor sequences. It is a complicated dense representations of channeled-spatio-temporal dynamics, containing values between 0 and 1. Low values show significant differences and weak reference qualities, whereas high values indicate strong correlative correlations and comparable patterns amongst the sensors. Let X^i_f be the sensor's i^{th} vector with f features at T^{th} time step. The MX score is thus calculated as follows:

$$MXscore(X)^f_{x,y} = MIC(X^f_x, X^f_y) \tag{1}$$

such that, $MXscore(X)^f_{x,y} \in [0, 1]$ indicates how much data from sensor x may be extrapolated to sensor y using the feature f and *MIC* is the *Maximum Information Coefficient*.

B. Traffic Network

The network of roads within a transportation system can be represented in the form of a directed graph $G = (V, E)$, where V represents a set of *vertices*, such that $N = |V|$. The vertices are considered as traffic nodes (sensors installed in the traffic network), and the connections between a set of vertices is represented by E, which marks a relation between the two connecting vertices.

C. Signal Matrix

The traffic observations for a *traffic network* G, at a given time p, is represented by a matrix defined as *'Signal Matrix'*, and is given as $Z^p = (z^{p_1}, z^{p_2}, \ldots, z^{p_N}) \in \mathbb{R}^{N\times F}$, where z^{p_N} is the feature vector of the N^{th} node and F is the number of features of node N at time p.

D. Problem Statement

Consider the case of a traffic node in the center of a traffic network. The traffic condition at this node could be connected to those in it's proximity and also to the distant nodes throughout the network, creating spatial relationships between the nodes. Additionally, there may be temporal relationships between the traffic lights at one node and those at the backward and forward nodes. Evidently, the traffic circumstances are node and time evolving much as spatial and temporal interactions. However, the signals are prone and non-resilient to surge in traffic collisions and other exceptional events that happen on the road. Therefore, it is challenging to exactly forecast how the increase in vehicle flow at one node, at one point of time, may affect the change in vehicle flows at other nodes at subsequent time steps due to these multi-aspect interactions and couplings. Existing works on self attention based consider that the module's input and output as two sequences of vectors and fail to achieve the desired performance. One of the reasons being the deployment of the self-attention module in the multi-dimensional data space due to the separate development of the module either in spatial or in the temporal dimension of the traffic data. Therefore, based on the previously collected spatio-temporal signal matrix:

$$Z = (z^{p-1}, z^{p-2}, \ldots, z^{p-x_t}) \in \mathbb{R}^{N\times F\times Z_{p-x_t}} \tag{2}$$

Our objective is to forecast the signal matrix as:

$$Y = (z^{p+1}, z^{p+2}, \ldots, z^{p+x_t}) \in \mathbb{R}^{N\times F\times Z_{p+x_t}} \tag{3}$$

TABLE I

COMPARATIVE ANALYSIS OF ADVANCED TRAFFIC PREDICTION MODELS

Models	**Complexity**	**Performance**	**Spatial Dependencies**		**Temporal Dependencies**		**Channel Dependencies**	**Advantages of Proposed Approach over Existing Methods**
			Static	**Dynamic**	**Short-term**	**Long-term**		
DCRNN [24]	Medium	Fair	✓	–	✓	✓	–	1- Improved accuracy due to weighted channel relations,
STGCN [18]	Medium	Fair	✓	–	✓	–	–	synchronous relation handling.
ASTGCN [13]	Medium	Good	✓	✓	✓	–	–	
WaveNet [25]	High	Fair	–	✓	✓	✓	–	2 - Scalable architecture due to light-head architecture
STSGCN [26]	High	Fair	✓	✓	–	✓	–	and a multi-dimensional self-attention scheme,
AGCRN [7]	High	Good	✓	–	✓	–	–	enabling parallel computation.
Traffic-GGNN [19]	Medium	Good	–	✓	✓	✓	–	
GAT-STC [20]	Medium	Good	–	✓	✓	✓	✓	3- Flexible data augmentation is achieved through multi-
STAGCN-EC [22]	Medium	Good	✓	✓	✓	–	–	component structure and time-gated fusion mechanism.
MSGAT [27]	High	Good	–	✓	✓	✓	✓	4- Multi-aspect coupling and multi-dimensional self-
STGAGRTN [21]	Medium	Good	✓	✓	✓	✓	–	attention scheme for robustness.
GSAA [23]	High	Good	✓	✓	✓	✓	–	
CCST-GAN (Proposed)	**Medium**	**Excellent**	✓	✓	✓	✓	✓	5 - Parameterization and parallel processing
								for optimized training and computation cost.

Such that, $(z^{p-1}, z^{p-2} \ldots\ldots z^{p-x_t})$ represent the time steps prior to z^p and $(z^{p+1}, z^{p+2}, \ldots, z^{p+x_t})$ represent the future time steps.

III. METHODOLOGY

The proposed CCST-GAN framework captures the inter-node interaction of traffic dynamics of network node's with each other in a three dimensional manner mainly:

1) Nodes spatio-temporal dynamics.
2) The inter-channel relationships of traffic signals and their impact on the dynamics of a road system.
3) The impact of these relationships on a single node's traffic flow.

The proposed CCST-GAN analyses the interrelated interactions between different features of traffic signals and incorporates a time-gated technique for traffic prediction. Using the multi-component fusion and the intricate channel, temporal and spatial connections between various traffic signal components, relevant properties across temporal sequences is identified. This information is then used to collect comparable patterns from every sensor and create a dynamic correlation information network, termed as the MXGCN. The proposed CCST-GAN integrates a robust data augmentation mechanism that leverages the entire temporal dimension of the available data to detect the latent patterns that are highly conducive to predictive modeling.

A. MX Score Based Prediction Unit

The CCST-GAN is a time-gated fusion based model which utilizes a complex multi-component structure and is made up of R identical *MX score based Predictor units (MXPUs)*, as shown in the Figure 1. For an accurate near-term prediction each *MXPU* component shares an identical network topology. Instead of relying on a heavy encoder-decoder framework, CCST-GAN adopts the light-head structural style, that has shown to be particularly effective in computer vision applications like target identification. Each *MXPU* is made up of many stacked *MX score based Embedding Modules (MXEMs)* and a light-head block connected to the final predictor, as shown in the Figure 1. The *MXEM* stacking replicas, in a *MXPU*, are in charge of extracting traffic dynamics from the input data to the present *MXPU* along with the complicated dynamic characteristics connected to the numerous relationships arising from the changing traffic circumstances. Deep prediction of the outcomes is achieved by regression mapping of the learned features by attaching a light-head block to it's prior *MXEM*. The light-head block, in this case, is sequentially made up of *Layer Normalization* and *Fully Connected Layer*, to resist the significant decoder overhead and guarantee the precision of short-term prediction. The n^{th} input and output of the *MXPUs*, such that $(1 \leq n \leq R)$, are formally referred as sample $X_i \in \mathbb{R}^{N \times F \times T_i}$ and $\hat{Y}_i \in \mathbb{R}^{N \times F \times T_o}$ respectively, where F is number of features associated with each node, N denotes the nodes in the traffic network, and T_i and T_o denote the input and output time-steps, respectively.

B. MX Score Based Graph Convolution Network

The proposed CCST-GAN uses the MXGCN to construct accurate features by associating the propagate-aggregate mode with the predefined graph structure and density similarity representation. The basic network, given by $s^{(\alpha)} = \sigma(As^{(\alpha-1)}W^{(\alpha)})$, suggests that the network only takes neighbor sensor connections into account, even if the neighbor sensors do not share characteristics and patterns. However, we suggest that the MXGCN module is capable of broadly catching comparable patterns from various sensor nodes and the revised formulation is given as:

$$\hat{s}^{(\alpha)} = AGG(\psi_f \sigma(MX_f h^{(\alpha-1)} W^{(\alpha)})) \tag{4}$$

where $MX_f \in \mathbb{R}^{TP \times TP}$ denotes the proposed MX score in dimension f, and ψ_f is the parameter to be trained.

Finally, using a trainable parameter 'ϕ', to prevent losing local information, a predefined graph structure is introduced and is defined as follows:

$$\hat{s}^{(\alpha)} = AGG(\hat{h}^{(\alpha)}, \phi s^{(\alpha)}) \tag{5}$$

In contrast to existing works, the MXGCN module combines structural information with MIC allowing it to provide the propagate-aggregate mode with precise and deep interconnections. The advantage is seen in terms of the proposed method creating more accurate features for other components. A steep proliferation in the feature aggregation efficiency is also seen across all the comparable sensors. The MXGCN is further used to feed spatial correlation information into the structural network.

C. MX Score Based Self-Attention Method

The proposed scheme makes use of the classical attention mechanism as illustrated in Figure 2. The mechanism

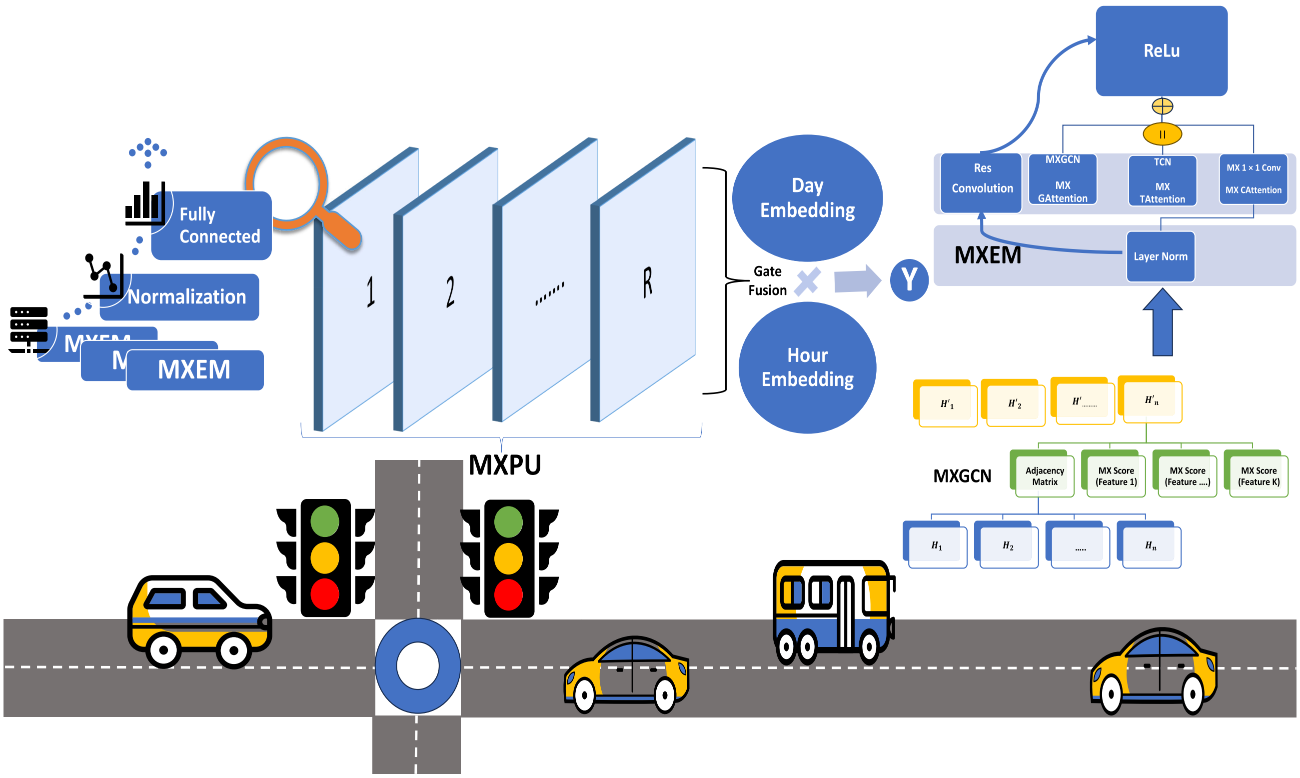

Fig. 1. The proposed framework.

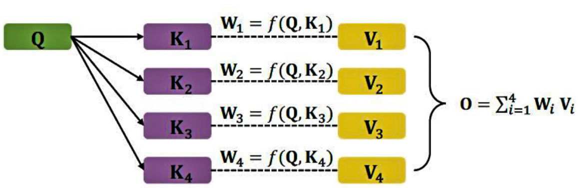

Fig. 2. Classical attention mechanism.

involves *query vectors*, *key-value pair vectors*, and *dot product vectors* of both as a result. Weighted sum of these values is considered as the final result, where weights indicate the relational strength between each key-value combination and query. The *classical self-attention mechanism*, also known as the *scaled dot-product attention mechanism*, calculates the attention weights by taking the dot product of the *query vectors* and *key vectors*, which is then scaled by the square root of the dimensionality of the *key vectors*, to prevent numerical instability [28]. This is used to determine the *attention score*, which is defined as the degree of correlation between the input vectors.

In the proposed CCST-GAN, we provide a sophisticated MIC based MX score for channel, spatial and temporal representations. This is used to assess the level of informational correlation between sensor sequences. To further understand the calculation of MIC, we define two sequences, $Z_1 \in \mathbb{R}^C$ and $Z_2 \in \mathbb{R}^D$. The information on the degree of correlation between Z_1 and Z_2 is calculated as shown in Equation 6.

$$MIC(Z_1, Z_2) = \frac{max_{C*D<C\delta}\{I_{c,d}(Z_1, Z_2)\}}{log_2 min\{C, D\}} \tag{6}$$

where δ is a parameter to control the partitions and $I_{c,d}$ denotes the mutual information between c and d and is calculated as:

$$I_{c,d} = \sum_{c<C,d<D} q(c,d) log_2(\frac{q(c,d)}{q(c)q(d)}) \tag{7}$$

where for the chosen grids $(c, d), q(c)$ and $q(d)$ represents the *Edge Probability Density* and $q(c, d)$ represents the *Joint Probability Density*. The input data sequence in traffic prediction tasks displays spatial characteristics that are interrelated. In contrast to *Dynamic Time Warping (DTW)* and *Cosine Similarity* approaches, MIC is able to quickly complete computations in addition to capturing a variety of relationships. With the goal of easily adapting the present traditional self-attention strategy to a multi-dimensional data space, we design a simple but effective multi-dimensional self-attention scheme.

Consider a $3D$ space with x, y, and z axes that transmit three unique forms of data connected to the original data. Specifically, the x, y, and z axes represent the *channel*, *geographical*, and *temporal*, aspects of channeled-spatio-temporal traffic signal data, respectively.

For simplicity's sake, the data space is designated as a discrete set $\{s_{x,y,z}\}$. If the information along the x axis has a possible correlation that needs to be discovered, the information along the y axis and z axis will cooperate to make that happen. Our self-attention scheme, in particular, replaces the vector used in conventional self-attention mechanisms with the same matrix represented as $(s_{x,:,:})$, such that, ' $:'$ represents elements in the associated dimensions. An efficient linear translation for tensor operations converts *query(Q)* and *key(K)*

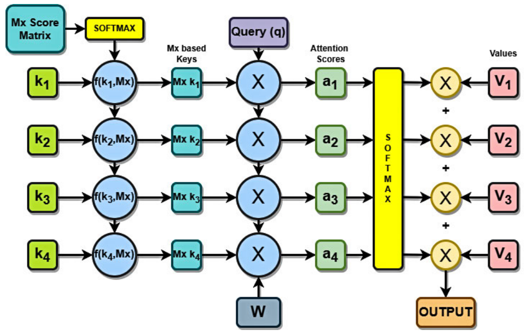

Fig. 3. MX score based self-attention scheme.

to vectors. Finally, the whole set of vectored keys and queries are then placed into the matrices Q and K. Based on this, the MX score based self-attention can be represented as:

$$A = Att(Q, \tilde{K}) = softmax(QW\tilde{K}^{\top}) \tag{8}$$

where W is a matrix with the dimensions $n_k \times n_k$ that represents all parameters that may be learned to apply an attention in x dimension, and n_k is the key used. Undoubtedly, having a 'n_k' that is too big, will make the application of the proposed attention method computationally expensive and will degrade the performance and utility of the final model. As such, to overcome the computational challenges brought by using too many parameters in the application of the proposed attention method, the matrix W is specifically substituted by product of a smaller learnable parameter matrix and it's transpose; i.e. $W = ww^{\top}$, where w denotes the smaller matrix with lower parameter values ($n_k \times w_k$) (in this case, $w_k < n_k$). The next step is to use the current attention weighted sums and update all values on the x axis by their hidden states synchronously. Figure 3 shows the working of the proposed *MX Score based Self-Attention Method*.

It is important to note that, the sensors with strong correlation information have similar traffic road network patterns. MX score is used to combine comparable sensor patterns to generate $\tilde{K}$ as:

$$\tilde{K} = \alpha(\mathrm{MX}, K) = \frac{1}{F}\sum_{F} \hat{\mathrm{MX}}, \hat{K} \tag{9}$$

where α is a reconstructive function. The inputs are the *MX score* and the *original key* and the output is $\tilde{K} \in \mathbb{R}^{N \times L_k \times d_k}$, where L_k is the length of k. This output is more steady for the self-attention mechanism and $\tilde{K}$ can be replaced to evaluate the Equation 8. In our approach, we combine the related sensor patterns from several sources with MX score weights to provide more reliable contextual characteristics for the attention mechanism. As a result, the attention mechanism acquires more concentrated attention weights (the same can be seen in the Figure 1). A time-series sequence may properly identify it's most relevant sequence with the use of MX score, which improves the accuracy of activities involving traffic flow forecasting.

D. MX Score Based Embedding Module (MXEM)

The MX score based embedding module (MXEM), serves as the foundational component of the *MXPU* and consists of three embedding branches that represent the *spatial*, *temporal*, and *channel* connections, respectively, and is shown in Figure 1. The deep characteristics associated with the complex channeled-spatio-temporal dynamics of traffic network are extracted by the *MXEM* stacking replicas. The embeddings collected from various branches are combined in the MXEM's, which uses its own self-attention mechanism to implicitly calculate each branch's relative relevance to the projected outcome.

1) Spatial Relation Embedding: The spatial relation embedding comprises of two crucial procedures; *MX GAttention* and *MXGCN*, and is shown in Figure 1. The *MX GAttention* operation is first applied to the MX score based self-attention method (suggested in III-C) to each MXEM's spatial dimension of incoming data. As an example, consider the α layer MXEM, in the i^{th} *MXPU* of the proposed CCST-GAN. The α represents the MXEM stacking replica identification number and it's input is indicated as $s^{(\alpha-1)} \in \mathbb{R}^{SP_\alpha \times TP_\alpha \times CH_\alpha}$, where $s^{(\alpha)} = x_i$, and SP_α, TP_α, and CH_α represent the spatio, temporal and channel dimensions of $s^{(\alpha-1)}$, respectively. Each spatial node has an identical matrix set as its *query*, *key*, and *value*, as defined in section. This is followed by encoding every query vector and key vector into their respective matrices, which are indicated as $Q_x \in \mathbb{R}^{TP_\alpha \times CH_\alpha}$ and $K_x \in \mathbb{R}^{TP_\alpha \times CH_\alpha}$, respectively. A new matrix $A_x^{(\alpha)} \in \mathbb{R}^{TP_\alpha \times TP_\alpha}$, showing the possible spatial relations, is created in accordance with equation 8. The *MX GAttention* operation is thus defined as:

$$\begin{aligned} MX\_GAtt(s^{(\alpha-1)}) &= A_x^{(\alpha)} = Att(Q_x, \tilde{K_x}) \\ &= softmax(Q_x W_x \tilde{K_x^{\top}}) \end{aligned} \tag{10}$$

We next use the *MXGCN* module, used in the previous layer's MXEM, to compile the hidden states of each node and their first-order neighbors $s^{(\alpha-1)}$. As shown in Figure 1, the MXGCN will feed spatial correlation information into the structural network. The basic network, used for this work is defined as:

$$s^{(\alpha)} = \sigma(As^{(\alpha-1)}W^{(\alpha)}) \tag{11}$$

where s^{α} and $s^{(\alpha-1)} \in \mathbb{R}^{TP \times d_k}$, $W^{(\alpha)} \in \mathbb{R}^{d_k \times d_k}$ and σ is the d_k-dimensional sensor feature representation of the output. α and $\alpha - 1$ denote the layer numbers and the Laplacian regularization of the normalized structural adjacency matrix yields $A \in \mathbb{R}^{TP \times TP}$, which is used as an adjacency matrix.

$$A = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}} \tag{12}$$

here $\tilde{D}$ is the diagonal matrix, and $\tilde{A} \in \mathbb{R}^{TP \times TP}$ is the adjacency matrix with the i^{th} element $\tilde{D}_{ii} = \sum A_{ij}$ of non-linear activation function.

2) Channel Relation Embedding: We construct another MXEM branch for connection embedding in channels as shown in Figure 1. *MX CAttention* inherits previous embedding branch procedures to detect pairwise channel correlations using the proposed MX self-attention mechanism. The information located in the other two dimensions is used together to compute the score for attention between various traffic channel dynamics. The correlation strength between channels is semantically represented by the attention score and the channel associated attention map, represented as $A_f^{(\alpha)} \in \mathbb{R}^{SP_\alpha \times SP_\alpha}$, is calculated by:

$$MX\_CAtt(S^{(\alpha-1)}) = A_f^{(\alpha)} = Attention(Q_f, \tilde{K_f}) = softmax(Q_f W_f \tilde{K_f}^\top) \quad (13)$$

where ($W_f \in \mathbb{R}^{CH_\alpha \times CH_\alpha}$) is the learnable parameter matrix, ($Q_f \in \mathbb{R}^{SP_\alpha \times CH_\alpha}$) and ($\tilde{K_f} \in \mathbb{R}^{SP_\alpha \times CH_\alpha}$) are the query matrix and MX score based key matrix, respectively. After dimension translation, the MX score based 1×1 convolution generates the channel embedding $e_f^{(\alpha)}$ followed by the current MXEM layer implicitly assessing it's relevant weight from previous MXEM layer, ensuring accurate prediction outcome.

3) Temporal Relation Embedding: In the actual road system, pair-wise traffic scenarios typically have different implicit or explicit temporal correlations at different time steps. For instance, the rate of traffic flow on the same traffic node may exhibit weekly or daily periodic patterns. One node's observation of the average vehicle speed, at one time step, may be connected to another nodes past traffic volume. To incorporate these temporal connections, we establish a new MXEM branch for temporal relation embedding as shown in Figure 1. The embedding branch includes the crucial procedures *MX TAttention* and *TCN*. The *MX TAttention* applies the proposed multi-dimensional self-attention mechanism to each MXEM's temporal input data. Considering the α layer of the MXEM in the i^{th} *MXPU* of the proposed CCST-GAN, the *MX TAttention* is defined as:

$$TAtt(s^{(\alpha-1)}) = Att(Q_T, K_T) = softmax(Q_T W_T K_T^\top) \quad (14)$$

where, ($Q_T \in \mathbb{R}^{CH_\alpha \times TP_\alpha}$) refers to the query matrix, ($K_T \in \mathbb{R}^{CH_j \times TP_j}$) refers to the key matrix and the parameter weight matrix is given by ($W_T \in \mathbb{R}^{TP_\alpha \times TP_\alpha}$). To further optimize the accuracy, *MX TAttention* is redefined as:

$$TAttention(s^{(\alpha-1)}) = Att(Q_T, K_T) = softmax(Q_T E_T E_T^\top K_T^\top) \quad (15)$$

where $E_T \in \mathbb{R}^{TP_\alpha \times d_E}$ and d_E is a smaller value than TP_α.

The acquired temporal attention map, given by ($A_T^{(\alpha)}$), is used to reset the temporal attention to ($s_T(\alpha-1) \in \mathbb{R}^{SP_\alpha \times TP_\alpha \times CH_\alpha}$), where $s_T(\alpha-1)$ is the embedding branch output and $s(\alpha-1)$ is the current MXEM input. To handle ($s_T^{(\alpha-1)}$) temporally, we use *TCN* which follows temporal CNN to maintain data order as current outputs are exclusively connected to historical data.

IV. Experiments and Results

Extensive trials were performed to validated the performance accuracy and efficiency of the proposed CCST-GAN. A detailed description of the experiments and the outcomes are presented below:

A. Data Description

For validating the performance of the proposed CCST-GAN and performing a comprehensive comparative analysis with the state-of-the-art methods, 5 public real time traffic network data sets were used. All the data sets are well established and are widely used in developing and validating solutions for accurate traffic prediction. The data sets are built using the Caltrans Performance Measurement System (PeMS) [29]. Each of the 5 data sets correspond to a different traffic route network and is related to a district in California. These data sets consolidate all time-series observations into 5 minute intervals, which results in 12 sample points (also known as time steps) per hour. The names and key features of all the five data sets, used in this study, are highlighted in Table II. It is important to mention that contrary to the *PEMSD3* and *PEMSD7*, which only includes traffic flow, the *PEMSD4*, *PEMSD8* and *PEMS-BAY*, have 3 different features associated with every node; *flow*, *occupancy* and *average speed*.

TABLE II
Details of the Data Set Used

Data Set name	Sensors or Nodes	Edges	Channels	Time Steps
PEMSD3 [30]	358	547	1	26208
PEMSD4 [30]	307	340	3	16992
PEMSD7 [30]	883	866	1	28224
PEMSD8 [30]	170	295	3	17856
PEMS-BAY [26]	325	8033	3	52116

On the basis of connection between network nodes and MX score, we created a spatial adjacency matrix for each dataset. We then used z-score normalization to uniformly normalize all the data. The 5 datasets were each subjected to our data augmentation strategy (discussed in Section III-D.1). Finally, the data items from each data set are divided into *validation*, *test* and *training* sets in the ratio of $2:2:6$.

B. Methods for Comparative Analysis

This work builds on the work presented in [27] and the technical parameters used in MSGAT [27] are replicated for the comparative analysis of the proposed CCST-GAN. The results of the existing methods are used as presented in [27], while the performance of the proposed CCST-GAN is added as a new value to the analysis. The methods used for comparative analysis are: DCRNN [24], STGCN [18], ASTGCN [13], WaveNet [25], STSGCN [26], AGCRN [7] and MSGAT [27].

C. Evaluating Metrics

The identified metrics for performance validation are: *Mean Absolute Error (MAE)*, *Mean Absolute Percentage Error (MAPE)*, and *Root Mean Square Error (RMSE)*. If $\mathcal{E}_{x,y}$ is the *ground truth*, and $\hat{\mathcal{E}}_{x,y}$ is the *predicted value* such that ($1 \le x \le N$), and ($1 \le y \le P$), where N represents the number of nodes in traffic network and P is the number of future time steps, the metrics are calculated as:

$$MAE = \frac{1}{NP} \sum_{x=1}^{N} \sum_{y=1}^{P} \left| \mathcal{E}_{x,y} - \hat{\mathcal{E}}_{x,y} \right| \quad (16)$$

$$MAPE = \frac{1}{NP} \sum_{x=1}^{N} \sum_{y=1}^{P} \left| \frac{\mathcal{Y}_{x,y} - \hat{\mathcal{E}}_{x,y}}{\mathcal{E}_{x,y}} \right| \times 100\% \quad (17)$$

TABLE III

ERROR COMPARISON OF CCST-GAN ON PEMSD DATA SET AND PEMS-BAY DATA SET

Data Sets	Metric	DCRNN [24]	STGCN [18]	ASTGCN [13]	WaveNet [25]	STSGCN [26]	AGCRN [7]	MSGAT [27]	**Proposed CCST-GAN**
PEMSD3	MAE	18.18 ± 0.15	17.49 ± 0.46	17.69 ± 1.43	19.85 ± 0.03	17.48 ± 0.15	16.13 ± 0.19	15.68 ± 0.24	**15.32 ± 0.13**
	MAPE(%)	18.91 ± 0.82	17.15 ± 0.45	19.40 ± 2.24	19.31 ± 0.49	16.78 ± 0.20	-	16.11 ± 0.13	**15.54 ± 0.11**
	RMSE	30.31 ± 0.25	30.12 ± 0.70	29.66 ± 1.68	32.94 ± 0.18	29.21 ± 0.56	28.42 ± 0.07	26.54 ± 0.16	**26.19 ± 0.15**
PEMSD4	MAE	24.70 ± 0.22	22.70 ± 0.64	22.93 ± 1.29	25.45 ± 0.03	21.19 ± 0.10	19.75 ± 0.11	19.54 ± 0.19	**19.10 ± 0.14**
	MAPE(%)	17.12 ± 0.37	14.59 ± 0.21	16.56 ± 1.36	17.29 ± 0.24	13.90 ± 0.05	13.00 ± 0.18	13.44 ± 0.28	**12.90 ± 0.16**
	RMSE	38.12 ± 0.26	35.55 ± 0.75	35.22 ± 1.90	39.70 ± 0.04	33.65 ± 0.20	32.41 ± 0.20	31.69 ± 0.15	**30.92 ± 0.13**
PEMSD7	MAE	25.30 ± 0.52	25.38 ± 0.49	28.05 ± 2.34	26.85 ± 0.05	24.26 ± 0.14	21.19 ± 0.09	20.47 ± 0.13	**20.26 ± 0.10**
	MAPE(%)	11.66 ± 0.33	11.08 ± 0.18	13.92 ± 1.65	12.12 ± 0.41	10.21 ± 1.65	8.95 ± 0.08	8.84 ± 0.14	**8.52 ± 0.12**
	RMSE	38.58 ± 0.70	38.78 ± 0.58	42.57 ± 3.31	42.78 ± 0.07	39.03 ± 0.27	35.12 ± 0.12	34.17 ± 0.08	**33.64 ± 0.10**
PEMSD8	MAE	17.86 ± 0.03	18.02 ± 0.14	18.61 ± 0.40	19.13 ± 0.08	17.13 ± 0.09	16.10 ± 0.11	14.78 ± 0.09	**14.29 ± 0.06**
	MAPE(%)	11.45 ± 0.03	11.40 ± 0.10	13.08 ± 1.00	12.68 ± 0.57	10.96 ± 0.07	10.27 ± 0.08	10.07 ± 0.12	**9.60 ± 0.14**
	RMSE	27.83 ± 0.05	27.83 ± 0.20	28.16 ± 0.48	31.05 ± 0.07	26.80 ± 0.18	25.62 ± 0.17	24.15 ± 0.07	**23.42 ± 0.09**

TABLE IV

ERROR COMPARISON OF CCST-GAN ON PEMS-BAY DATA SET FOR BASELINE MODELS

Time Interval	Metric	DCRNN [24]	STGCN [18]	ASTGCN [13]	WaveNet [25]	STSGCN [26]	AGCRN [7]	MSGAT [27]	**Proposed CCST-GAN**
15 min	MAE	1.38	1.36	1.20	1.30	2.54	1.16	1.13 ± 0.02	**1.09 ± 0.03**
	MAPE(%)	2.90	2.90	2.34	2.73	5.88	2.47	2.44 ± 0.06	**2.37 ± 0.05**
	RMSE	2.95	2.96	2.43	2.74	4.79	2.40	2.38 ± 0.04	**2.35 ± 0.04**
30 min	MAE	1.74	1.81	1.46	1.63	2.60	1.41	1.35 ± 0.01	**1.27 ± 0.02**
	MAPE(%)	3.90	4.17	3.09	3.67	6.03	3.12	3.09 ± 0.04	**3.03 ± 0.03**
	RMSE	3.97	4.27	3.27	3.70	4.93	3.10	2.94 ± 0.02	**2.81 ± 0.03**
60 min	MAE	2.07	2.49	1.83	1.95	2.71	1.79	1.74 ± 0.03	**1.69 ± 0.04**
	MAPE(%)	4.90	5.79	4.15	4.63	6.39	4.01	3.97 ± 0.05	**3.89 ± 0.04**
	RMSE	4.74	5.69	3.20	4.52	5.28	3.78	3.88 ± 0.02	**3.10 ± 0.03**

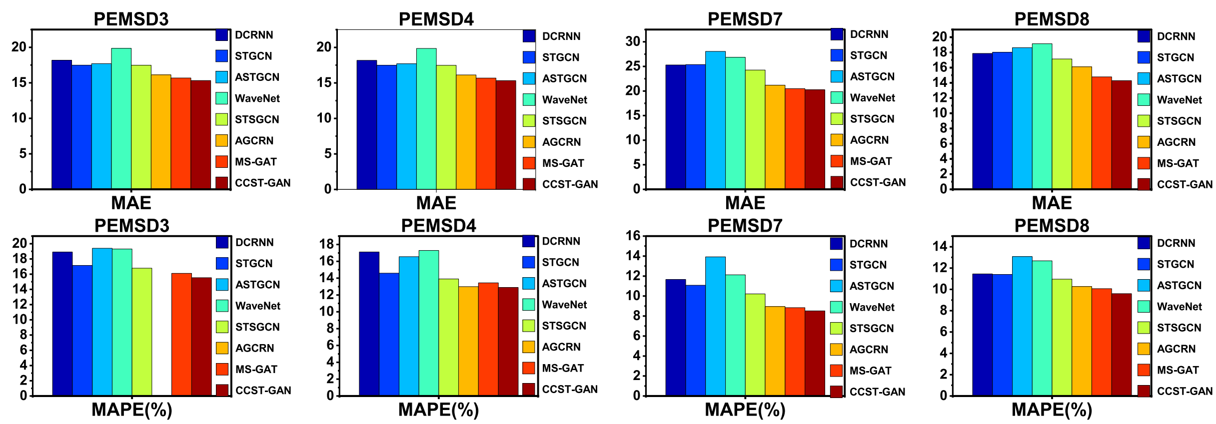

Fig. 4. MAE and MAPE of CCST-GAN on PEMSD3, PEMSD4, PEMSD7 and PEMSD8.

$$RMSE = \sqrt{\frac{1}{NP}\sum_{x=1}^{N}\sum_{y=1}^{P}\left(\mathcal{E}_{x,y} - \hat{\mathcal{E}}_{x,y}\right)^2} \tag{18}$$

D. Implementation Parameters

CCST-GAN accepts input as a multi-segment historical data proportional to the number of components used in prediction (i.e., MXPUs). The input samples have fixed time steps that are set to 12 for each segment. As such, CCST-GAN assigns a separate prediction component to each channeled-saptio-temporal sample collected in an hour before the forecast time period. This hour may be the current or the previous, from the day or week before the time period being used for forecasting. The models were used to predict traffic conditions for 1 *hour*, 30 *minutes*, and 25 *minutes* into the future. For all the models, the sensitivity of squared error loss is controlled by the hyper-parameter and is set at 40. Additionally, for all the models, *Adam* is used as an optimizer with a learning rate of $\frac{1}{1000}$ to train for a training epoch of 150 and the average results for 50 iterations are presented.

$$\mathcal{H}_\delta(\mathcal{X}, \hat{\mathcal{X}}) = \begin{cases} \frac{1}{2}\left(\mathcal{X} - \hat{\mathcal{X}}\right)^2 & if \ \left|\mathcal{X} - \hat{\mathcal{X}}\right| \le \delta \\ \delta\left|\mathcal{X} - \hat{\mathcal{X}}\right| - \frac{1}{2}\delta^2 & otherwise \end{cases} \tag{19}$$

E. Results and Analysis

The performance of all the models on PEMSD data set and PEMS-BAY data set is presented in Table III and Table IV, respectively. Table III and Figure 4 show that the proposed CCST-GAN is able to regularly outperform the compared methods on all the versions of the PEMSD data sets. Furthermore, Table IV and Figure 5 validate the performance of the proposed CCST-GAN as it is able to achieve improved accuracy on the PEMS-BAY data set as well. Results meet the expectations of the proposed CCST-GAN to be effective for short-term as well as for long-term predictions. Through further comparison, the following observations are made:

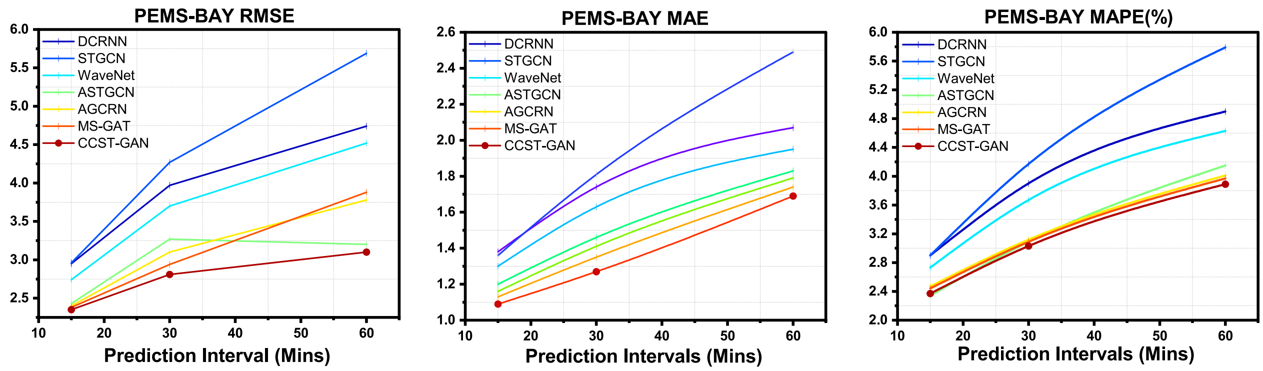

Fig. 5. MAE, MAPE and RMSE of CCST-GAN on PEMS-BAY at different time intervals.

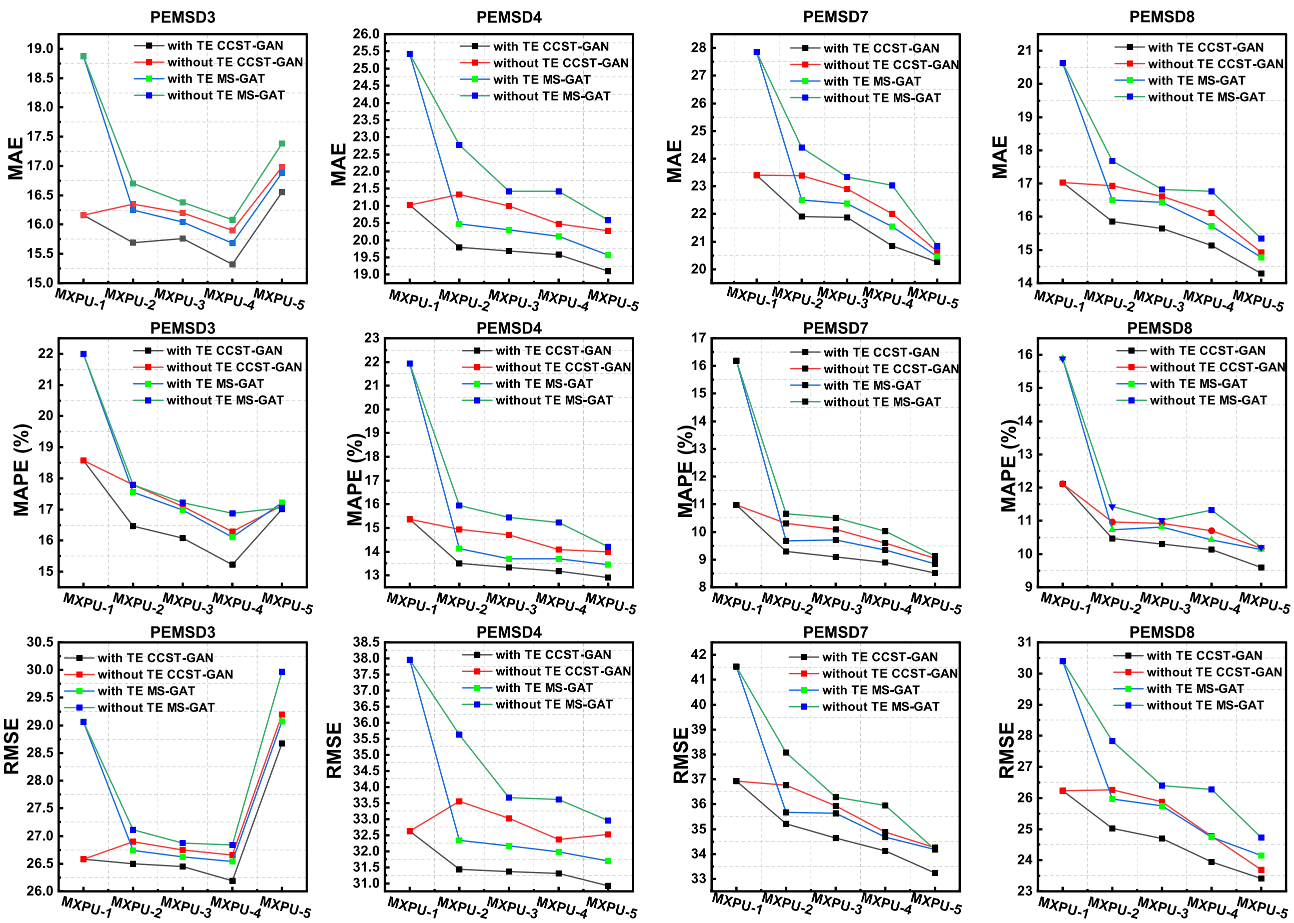

Fig. 6. Sensitivity analysis on PEMSD data sets with different number of MXPU units.

1) Existing methods of graph modeling for traffic predictions are primarily driven by saptio-temporal dynamics. However, the techniques of mapping the geographical and temporal linkages is unable to generate high performance output and are constrained by the frequently changing spatio-temporal relationships. The proposed CCST-GAN is able to address this issue by gaining knowledge about the couplings between geographical, channel and temporal connections and their impact on future traffic situations. It captures the traffic time-series linkages and the synchronous relation modeling allows optimized performance. The proposed CCST-GAN dynamically distributes the important weight of spatial, temporal, and channel linkages via adaptive learning from the ground-truth, making the regression model more predictive.
2) Identifying the importance of the model architecture for improved traffic prediction performance, the proposed CCST-GAN uses a multi-component architecture and is able to outperform the existing methods with a fair margin. Analyzing the performance of the methods, we identify that ASTGCN [13] uses *ChebNet* operation, as a graph convolution to describe spatial correlations which makes it a weaker predictive model. The method integrates input at a nearby time slice by stacking the temporal dimension in a conventional convolution layer,

similar to how WaveNet [25] does, but this makes it more challenging to produce a promising ability of modeling temporal correlation with dilated casual convolution. Additionally, it creates a highly complex spatio-temporal attention mechanism that captures the spatio-temporal dynamic correlations in the network. The effect of this is seen in large number of parameters for the model, making training difficult and prone to over-fitting. In contrast, CCST-GAN addresses these deficiencies by building a unique multi-component time-gating architecture and concentrating on the deficient characteristics.
3) The CCST-GAN considerably decreases the difficulty of aggregating spatial information and approximates the impact of conventional *ChebNet* of higher order, by stacking numerous GCN operations in MXEMs. For the purpose of temporal information aggregation, the proposed CCST-GAN leverages Temporal Convolutional Networks (TCNs) to take the advantage of the casual convolution and a more versatile receptive field in sequence modeling. The proposed method is able to capture complicated temporal traffic dynamics, opposed to employing CNN or RNN-based approaches in the past, such as the STGCN [18]).
4) The CCST-GAN highlights the relevance of the channel and presents a simple co-related self-attention method which is applicable to all the information dynamics of traffic circumstances. The proposed CCST-GAN uses the MXGCN to construct accurate features by associating the propagate-aggregate mode with the predefined graph structure and density similarity representation. Additionally, the proposed CCST-GAN assigns weight to the actual significance of the input graphs of the time-series data for each component, generating a more varied output sequence and, in turn, stabilizing the contextual features.
5) Analyzing the computational complexities of all the models, including the proposed CCST-GAN, it is observed that the proposed method uses graph convolutions with complexity of $O(E \cdot F \cdot F' + N^2 \cdot TC + T^2 \cdot NC)$. This is a significant improvement in terms of maintaining a balance between the prediction accuracy and computational complexity. Comparing the other existing works, such as the MSGAT [27], DCRNN [24] and STGCN [18], the complexities of these methods are computed to be fairly high at $O(E \cdot F \cdot F')$ and extends with multi-dimensional self-attention mechanisms to $O(N^2 \cdot TC + T^2 \cdot NC + C^2 \cdot NT)$, $O(E \cdot F \cdot F' + T \cdot N \cdot F^2)$ and $O(E \cdot F \cdot F' + T \cdot N \cdot F \cdot K)$, respectively. Further, ASTGCN [13] employs both graph convolutions and attention mechanisms, with complexities $O(N^2 \cdot F + T \cdot N \cdot F \cdot K)$, focusing dynamically on important features. WaveNet [25] utilizes adaptive graph convolutions and dilated causal convolutions, yielding $O(N^2 \cdot F \cdot F' + T \cdot N \cdot F \cdot K)$, capturing spatial-temporal dependencies through adaptive and dilated mechanisms. Finally, AGCRN [7], with adaptive graph convolutions and recurrent networks, has complexity of $O(N^2 \cdot F \cdot F' + T \cdot N \cdot F^2)$, adapting dynamically to changing graph structures, which leads to a comparatively higher computational demands.

F. The Ablation Study

The ablation study is presented for the proposed CCST-GAN on the 4 PEMSD data sets used in this study. Additionally, a comparative analysis of the outcomes of the sensitivity analysis, on the PEMSD data sets with different number of MXPU units, is also presented (Figure 6). The performance is validated against the performance of the MSGAT [27] method and the key observations are highlighted below:

1) Comparing the performance, it is evident that the proposed CCST-GAN is able to outperform the compared methods with a significant margin.
2) The effectiveness of the structured multi-component model for the channeled-spatio-temporal prediction is observed. Adding extra *MXPU* components to the CCST-GAN significantly improves the prediction accuracy.
3) Temporal Embedding (*TE*) is observed to be crucial for ensuring the model's correctness. The results in Figure 6 present the CCST-GAN's performance with and without *TE*. Improved performance is observed with *TE*, revealing that the traffic conditions present at separate time intervals in the past, have varying correlation values with those at the same time intervals in the future.

V. Conclusion

This work presents a novel co-related channeled-spatio-temporal graph attention method, termed as CCST-GAN, for accurate real-time traffic prediction in ITS. Analyzing the interrelated interactions between different features of traffic signals and corresponding correlation structure, we first propose a MX score based channeled-spatio-temporal representations for sensor sequences. The MX score is then used to design three crucial components called *MXPU*, *MXGCN* and *MXEM*. The MXPU and MXGCN are used to increase the efficiency of feature aggregation and generate more accurate features. To extract features from pertinent sequences, the MXEM is used to generate more targeted attention weights. The correlations between various periodic data is extracted and utilized to devise an accurate prediction solution for periodic data sets. The performance is validated on different versions of the publicly available PEMS-BAY and the PEMSD data set. The results demonstrate that in terms of predictive performance, the proposed CCST-GAN outperforms the existing state-of-the-art methods with a fair margin.

References

[1] L. Cao, "Non-IID recommender systems: A review and framework of recommendation paradigm shifting," *Engineering*, vol. 2, no. 2, pp. 212–224, Jun. 2016.
[2] C. Zhu, L. Cao, and J. Yin, "Unsupervised heterogeneous coupling learning for categorical representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 533–549, Jan. 2022.

[3] E. Cascetta, *Transportation Systems Engineering: Theory and Methods*, vol. 49. Cham, Switzerland: Springer, 2013.
[4] E. Zivot and J. Wang, "Vector autoregressive models for multivariate time series," in *Modeling Financial Time Series With S-Plus®*. New York, NY, USA: Springer, 2006, pp. 385–429. [Online]. Available: https://doi.org/10.1007/978-0-387-32348-0_11
[5] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for short-term traffic forecast," *IET Intell. Transp. Syst.*, vol. 11, no. 2, pp. 68–75, Mar. 2017.
[6] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. C, Emerg. Technol.*, vol. 54, pp. 187–197, May 2015.
[7] L. Bai, L. Yao, C. Li, and X. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," in *Proc. NIPS*, 2020, pp. 17804–17815.
[8] Z. Pan et al., "Spatio-temporal meta learning for urban traffic prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 3, pp. 1462–1476, Mar. 2022.
[9] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] C. Zheng, X. Fan, C. Wang, and J. Qi, "GMAN: A graph multi-attention network for traffic prediction," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 1, pp. 1234–1241.
[12] Y. Liang, S. Ke, J. Zhang, X. Yi, and Y. Zheng, "GeoMAN: Multi-level attention networks for geo-sensory time series prediction," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3428–3434.
[13] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial–temporal graph convolutional networks for traffic flow forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 922–929.
[14] Y. Wu and H. Tan, "Short-term traffic flow forecasting with spatial–temporal correlation in a hybrid deep learning framework," 2016, *arXiv:1612.01022*.
[15] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, p. 818, 2017.
[16] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2017, vol. 31, no. 1, pp. 1655–1661. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10735
[17] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
[18] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," 2017, *arXiv:1709.04875*.
[19] Y. Wang, J. Zheng, Y. Du, C. Huang, and P. Li, "Traffic-GGNN: Predicting traffic flow via attentional spatial–temporal gated graph neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18423–18432, Oct. 2022.
[20] Y. Chen et al., "Graph attention network with spatial–temporal clustering for traffic flow forecasting in intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 8, pp. 8727–8737, Aug. 2022.
[21] D. Wu, K. Peng, S. Wang, and V. C. M. Leung, "Spatial–temporal graph attention gated recurrent transformer network for traffic flow forecasting," *IEEE Internet Things J.*, vol. 11, no. 8, pp. 14267–14281, Aug. 2023.
[22] Q. Lai, J. Tian, W. Wang, and X. Hu, "Spatial–temporal attention graph convolution network on edge cloud for traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 4565–4576, Apr. 2023.
[23] J. Liu, X. Wang, H. Lin, and F. Yu, "GSAA: A novel graph spatiotemporal attention algorithm for smart city traffic prediction," *ACM Trans. Sensor Netw.*, Nov. 2023. [Online]. Available: https://doi.org/10.1145/3631608
[24] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," 2017, *arXiv:1707.01926*.
[25] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph WaveNet for deep spatial–temporal graph modeling," 2019, *arXiv:1906.00121*.
[26] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial–temporal synchronous graph convolutional networks: A new framework for spatial–temporal network data forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, 2020, pp. 914–921.
[27] J. Huang, K. Luo, L. Cao, Y. Wen, and S. Zhong, "Learning multi-aspect traffic couplings by multirelational graph attention networks for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20681–20695, Nov. 2022.
[28] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.
[29] C. Chen, K. Petty, A. Skabardonis, P. Varaiya, and Z. Jia, "Freeway performance measurement system: Mining loop detector data," *Transp. Res. Record, J. Transp. Res. Board*, vol. 1748, no. 1, pp. 96–102, Jan. 2001.
[30] H. Mao, "PEMS," *IEEE Dataport*, 2022. [Online]. Available: https://dx.doi.org/10.21227/fvne-xc16

Vishal Krishna Singh received the bachelor's degree in information technology, the master's degree in computer technology and application, and the Ph.D. degree in information technology from Indian Institute of Information Technology Allahabad, Allahabad, India, in 2018. He is currently a Lecturer and associated with the Networks and Communications Research Group, School of Computer Science and Electronics Engineering, University of Essex, Colchester, U.K. His research interests include the Internet of Things, wireless sensor networks, in-network inference, machine learning, and data analytics.

Neeraj Jain received the M.Tech. and Ph.D. degrees from Indian Institute of Information Technology Allahabad, India. He is currently an Assistant Professor with the Computer Science and Engineering and Information Technology Department, Jaypee Institute of Information Technology, Noida, India. His research interests include federated learning, machine learning, manifold learning, Internet of Things, and localization protocols.

Gaurav Tripathi (Member, IEEE) received the Ph.D. degree in computer science from the Department of Computer Science, Indian Institute of Information Technology Lucknow, Lucknow, India. He is currently a Data Scientist at GE Vernova. His research interests include machine learning, the Internet of Things, deep learning, and AI governance.

Sharad Sahani received the bachelor's degree in computer science and engineering (demonstrating early promise in the field of technology) and the master's degree in computer science (further honing his expertise) from the International Institute of Information Technology (IIIT) Lucknow, India, in 2019 and 2023, respectively. He is currently an integral part of AMD's Central Methodology Team, where he applies his knowledge to drive innovation and efficiency. His research interests encompass machine learning, the Internet of Things (IoT), and data analytics.