Dual-Scale Attributed Graph Transformer for Extracting Spatial-Temporal Features With Applications in Quality Index Prediction

Kesheng Zhang, Wen Yu, *Senior Member, IEEE*, and Tianyou Chai, *Life Fellow, IEEE*

***Abstract*—This paper presents a novel deep learning architecture, the Dual-scale Attribute Graph Transformer (DAGT), for extracting spatial-temporal features from attributed graph data. DAGT addresses the challenge of inconsistent sampling periods in industrial data streams by utilizing two key modules: 1) Dual-Scale Spatial-temporal Graph Convolution Network (DSGCN): This module captures both spatial and temporal information within attributed graphs, enabling effective feature extraction for tasks like quality index prediction. 2) Spatial-temporal Graph Attention Block (SGAB): This module employs an attention mechanism to selectively focus on crucial areas of the graph sequence. By assigning higher weights to regions with significant spatial-temporal features, SGAB refines the feature representation. The contributions of DAGT lie in the construction of a dual-scale adjacency matrix for efficient temporal and spatial dimensionality reduction and the design of a graph pooling module via spatial clustering. These innovations enhance the model's ability to learn from attributed graph sequences. The proposed method for quality index prediction is validated using real-world industrial data of the mineral processing process and various comparative experiments.**

***Index Terms*—Attributed graph, dual-scale, graph convolution networks, spatial-temporal features, quality index.**

I. Introduction

QUALITY index is a crucial parameter formulating the industrial processes and guiding production. Generally speaking, the production goal of industrial enterprises is to increase productivity while ensuring that the quality indices are within the target range. Multiple sub-processes in the industrial process need close cooperation and collaborative production control to ensure that the quality index is qualified. The quality index is mainly obtained by off-line sampling and laboratory testing. The production operators adjust the operational indices and production control set values of the process control system according to the quality index real value. The lag of off-line assay of quality index in industrial processes further hinders the optimization of production operation and stability control.

With the rapid advancement of the new generation of information and communication technology, the time series data of changing operating conditions accumulated in industrial production contain rich information about process knowledge, operational experience, and product quality. Mining the implicit relationship between high-frequency dynamic industrial data and the low-frequency offline assay quality index holds essential value for process control and operational decision-making [1], [2], [3]. Quality index prediction has become an important research domain of industrial intelligent manufacturing [4].

Due to the complexity and unclear mechanism of industrial processes, most scholars use data-driven modeling methods for industrial indices prediction. Some common data-driven time series prediction networks include Long Short-Term Memory (LSTM) [5], Convolutional Neural Network (CNN) [6], Gate Recurrent Unit (GRU) [7], or their deformations and stack structures [3], [8], [9], [10], [11]. However, a detailed analysis of the spatial-temporal features of industrial processes is lacking in the above researches. There are complex interactions, such as delay correlation and spatial correlation between input variables and between input and output variables. Therefore, these time series prediction networks struggle to fully capture the hierarchical causal relationships between process data and quality indices within structured space.

In the last five years, graph neural networks (GNNs) have achieved significant advancements in traffic forecasting [12], time series forecasting [13], [14], image recognition [15], and other fields. Powerful methods such as Graph convolution networks (GCNs) [16] and its variants have been widely used in these spatial-temporal network data prediction tasks [17], [18], [19], [20], and achieved good performance. They have also garnered significant attention and response in the industrial field [21], [22], [23], hoping to provide a new solution for capturing the spatial-temporal characteristics of industrial time series data. Existing GCNs or their approximations require explicit graph structures to realize the modeling. Unlike the image or transportation field, industrial multivariate time series

Received 23 May 2024; revised 29 July 2024; accepted 24 August 2024. This work was supported in part by the Key Research and Development Program of Liaoning Province under Grant 2023JH26 and Grant 10200011, in part by the Research Program of the Liaoning Liaohe Laboratory under Grant LLL23ZZ-05-01, and in part by the National Natural Science Foundation of China under Grant 61991404. *(Corresponding authors: Tianyou Chai; Wen Yu.)*

Kesheng Zhang and Tianyou Chai are with the Key Laboratory of Integrated Automation for Process Industry, Northeastern University, Shenyang 110004, China (e-mail: zks0053@163.com; tychai@mail.neu.edu.cn).

Wen Yu is with the Departamento de Control Automatico, CINVESTAVIPN (National Polytechnic Institute), Mexico 07360, Mexico (e-mail: yuw@ctrl.cinvestav.mx).

Recommended for acceptance by Y. Wang.

Digital Object Identifier 10.1109/TETCI.2024.3462486

data rarely reveal specific graph knowledge. Li et al. [21] used the graph learning algorithm concerning Granger causality to guide directional edge connection of the hierarchy to construct the adjacency matrix and realize time series prediction in the industrial process. Song et al. [22] proposed a graph construction block based on the maximum information coefficient (MIC) and developed a wind power prediction network using GCNs and multiresolution convolutional neural networks. However, most industrial time series data are located in structured spaces due to the intricate spatial-temporal correlation between the quality index and dual-scale industrial data, which may vary over time and operating conditions. These methods rely heavily on prior knowledge and perform poorly in the adaptability of dynamic spatial-temporal features mining of different variables in different production cells.

For the above problems, this paper proposes a Temporal Attributed Graph Transformer (DAGT) based on industrial dual-scale attributed graph, which extracts spatial-temporal characteristic from high and low-frequency industrial attributed graph data to predict quality index. The key contributions are as follows:

- Dual-Scale Attributed Graph Transformer (DAGT): A novel deep learning architecture is proposed to extract spatial-temporal features from attributed graph data. This framework addresses the challenge of inconsistent sampling frequencies often encountered in industrial data streams.
- Dual-Scale Spatial-temporal Graph Convolution Network (DSGCN): It is designed to capture both spatial and temporal information within attributed graphs. It facilitates effective feature extraction for tasks requiring an understanding of spatial-temporal relationships. The construction of a dual-scale adjacency matrix is a critical aspect of DSGCN, allowing for efficient dimensionality reduction in the learning process.
- Graph Pooling Module via Spatial Clustering: This paper introduces a novel graph pooling module that uses variables spatial clustering in industrial processes. This module reduces the complexity of the graph data while preserving important spatial information, improving the overall efficiency of DAGT.
- Spatial-temporal Graph Attention Block (SGAB): This module utilizes an attention mechanism, which selectively focuses on crucial areas of the graph sequence, assigning higher weights to regions containing significant spatial-temporal features.

II. Problem Statement

There are two key challenges in extracting meaningful features for quality index prediction in industrial processes:

1) The inherent nature of the data itself poses difficulties. Industrial data is often structured as attributed graphs, where nodes represent components and edges capture their relationships. However, these relationships and the associated data (attributes) are not static – they evolve over time, introducing crucial spatial-temporal features.

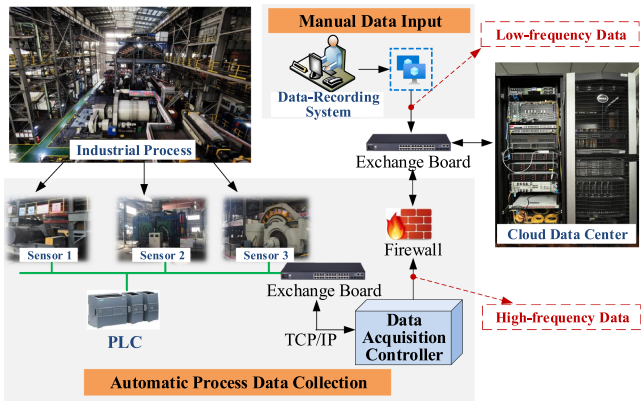

Fig. 1. Flow chart of data collection in industrial processes.

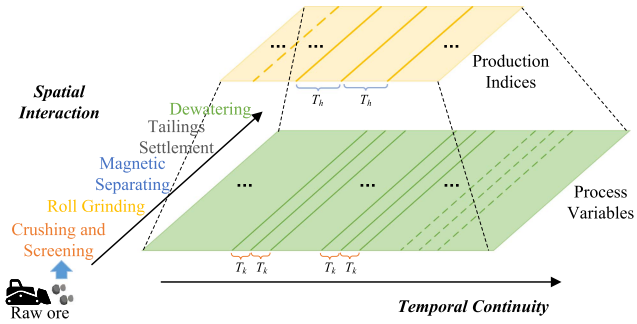

Fig. 2. Spatial-temporal relationship between process variables and production indices across the entire production line.

2) The way data is collected presents another hurdle. Inconsistent sampling frequencies disrupt the ability to capture these temporal aspects effectively. Sensor readings, for instance, might be taken at irregular intervals, leading to gaps in the data that traditional methods struggle to handle.

In this section, the quality index prediction and the dual-scale spatial-temporal characteristic of the quality index are discussed and analyzed to show the key challenges.

A. Quality Index in Industrial Processes

Industrial processes have many production parameters, which are coupled with each other and have strong nonlinear relationships. The flow chart of data collection in industrial processes is shown in Fig. 1. The relevant indices of the industrial process are divided into two categories: process variables with T_x sampling period and production indices with T_p sampling period. The spatial-temporal relationship between process variables and production indices across the entire production line is shown in Fig. 2. Usually, the sampling period of process variables is much faster than that of production indices. The details are as follows [24]:

- *Process Variable:* It refers to the high-frequency production process data directly collected by the Data Acquisition Controller (DAC) through various sensors or machinery, such as equipment motor current, slurry, liquid

level, fluid flow, etc. Process variables exhibit characteristics such as being multivariate, having high-frequency variability, and demonstrating strong fluctuations and uncertainty.

- *Production Index:* It reflects the quality, efficiency, consumption, and other relevant, comprehensive indices of the whole production line or the processing unit of process industries, such as quality index, yield index, energy consumption index, etc. More specifically, the production indices can be subdivided into comprehensive production indices and operational indices according to the production and operational decision-making process.

B. Quality Index Prediction

Assuming that the original process variables at time k is $\boldsymbol{X}(k) = [\boldsymbol{x}_1(k), \boldsymbol{x}_2(k), \ldots, \boldsymbol{x}_N(k)]^{\mathrm{T}}$ and the original production indices at time k is $P(k) = [p_1(k), p_2(k), \ldots, p_{N_L}(k)]^{\mathrm{T}}$. N is the number of process variables,while N_L is the number of production indices. The quality index at time k, denoted as $y(k)$, typically shares the same sampling period as the production indices. The dual-scale gap is defined as $\Gamma = T_x/T_p$, which is also the length of process variables data obtained by sampling at T_x interval in the sampling period T_p.

Within a T_p production cycle, the production indices are constructed into a one-dimensional vector $P(k) \in \mathbb{R}^{N_L}$, while the process variables with T_x sampling period are constructed into a two-dimensional matrix $\boldsymbol{X}(k)$, which is expressed by (1).

$$\begin{aligned}\boldsymbol{X}(k) &= [\boldsymbol{x}_1(k), \boldsymbol{x}_2(k), \ldots, \boldsymbol{x}_N(k)]^{\mathrm{T}} \\ &= \begin{bmatrix} x_{1,1}^k, & x_{1,2}^k, & \cdots, & x_{1,N}^k \\ x_{2,1}^k, & x_{2,2}^k, & \cdots, & x_{2,N}^k \\ \vdots & \vdots & \ddots & \vdots \\ x_{\Gamma-1,1}^k, & x_{\Gamma-1,2}^k, & \cdots, & x_{\Gamma-1,N}^k \\ x_{\Gamma,1}^k, & x_{\Gamma,2}^k, & \cdots, & x_{\Gamma,N}^k \end{bmatrix}^{\mathrm{T}} \end{aligned} \tag{1}$$

where $\boldsymbol{X}(k) \in \mathbb{R}^{N\times\Gamma}$, k is the time sequence index, $x_{\Gamma,N}^k$ is the Γ^{th} sampling period value of the N^{th} process variable in the k^{th} production period.

There is an unknown high-order nonlinear relationship between quality index and operational indices, key production statistical indices, and production process variables, making it difficult to accurately establish a quality index prediction model based on mechanism and prior knowledge. In a dynamic industrial environment, the multi-step prediction model for quality index is a dynamic nonlinear system with unknown structure, parameters, and order, as shown in (2).

$$\begin{aligned} y(k+\tau_y) = g(&\boldsymbol{X}(k), \boldsymbol{X}(k-1), \ldots, \boldsymbol{X}(k-\sigma_x+1), \\ &P(k), P(k-1), \ldots, P(k-\sigma_p+1)) \end{aligned} \tag{2}$$

where $g(\cdot)$ is the unknown nonlinear model, τ_y is prediction step of quality index, σ_x, σ_p are model orders for predicting quality index influenced by process variables and production indices.

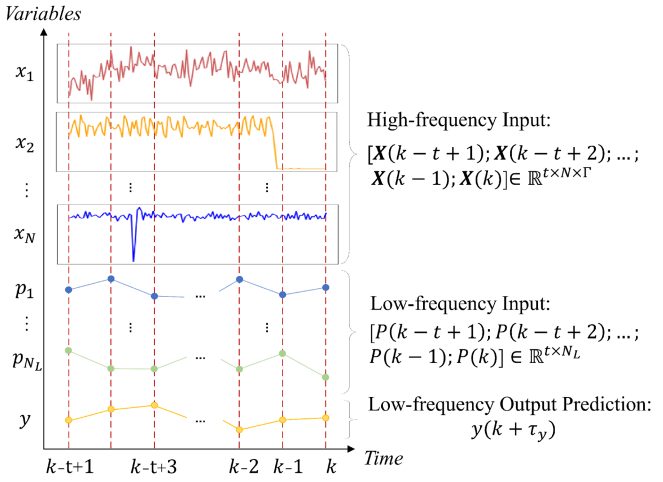

Fig. 3. The temporal expansion process of dual-scale industrial data.

C. Dual-Scale Spatial-Temporal Characteristic of Quality Index

Assuming that the model order for predicting quality index is t ($\sigma_x = \sigma_p = t$), the sliding time window technique is employed to extend the dual-scale industrial data and construct the high-frequency time series matrix $\boldsymbol{D}_H(k) \in \mathbb{R}^{t\times N\times\Gamma}$ and the low-frequency time series matrix $\boldsymbol{D}_L(k) \in \mathbb{R}^{t\times N_L}$, as shown in (3). The temporal expansion process of dual-scale industrial data is shown in Fig. 3.

$$\begin{aligned} \boldsymbol{D}_H(k) &= [\boldsymbol{X}(k-t+1); \boldsymbol{X}(k-t+2); \ldots; \boldsymbol{X}(k)] \\ \boldsymbol{D}_L(k) &= [P(k-t+1); P(k-t+2); \ldots; P(k)] \end{aligned} \tag{3}$$

The dual-scale signal $[\boldsymbol{D}_H(k), \boldsymbol{D}_L(k)]$ is constructed by (3). Then, (2) can be rewritten as (4):

$$y(k+\tau_y) = g(\boldsymbol{D}_H(k), \boldsymbol{D}_L(k)) \tag{4}$$

Dual-scale industrial data refers to datasets that encompass information collected from two distinct temporal resolutions or sampling frequencies within an industrial context. The spatial distribution position of variables and the high and low-frequency time series sampling reflect the spatial-temporal features of industrial data. The dual-scale spatial-temporal characteristics include dual-scale characteristics, temporal continuity, and spatial interactivity. The details are as follows:

- *Dual-scale Characteristic:* The dual-scale industrial data comprises information collected from two different temporal resolutions or scales within an industrial setting. This may include high-frequency data, capturing rapid changes and fluctuations occurring over short intervals (e.g., seconds or minutes), and low-frequency data, representing slower trends and patterns observed over longer intervals (e.g., hours, shifts, or days). Dual-scale characteristic presents both challenges and opportunities for modeling and prediction tasks in industrial applications. Models need to effectively capture dependencies and interactions between variables at different temporal scales.
- *Temporal Continuity:* It refers to the uninterrupted temporal stream of industrial data in industrial processes over

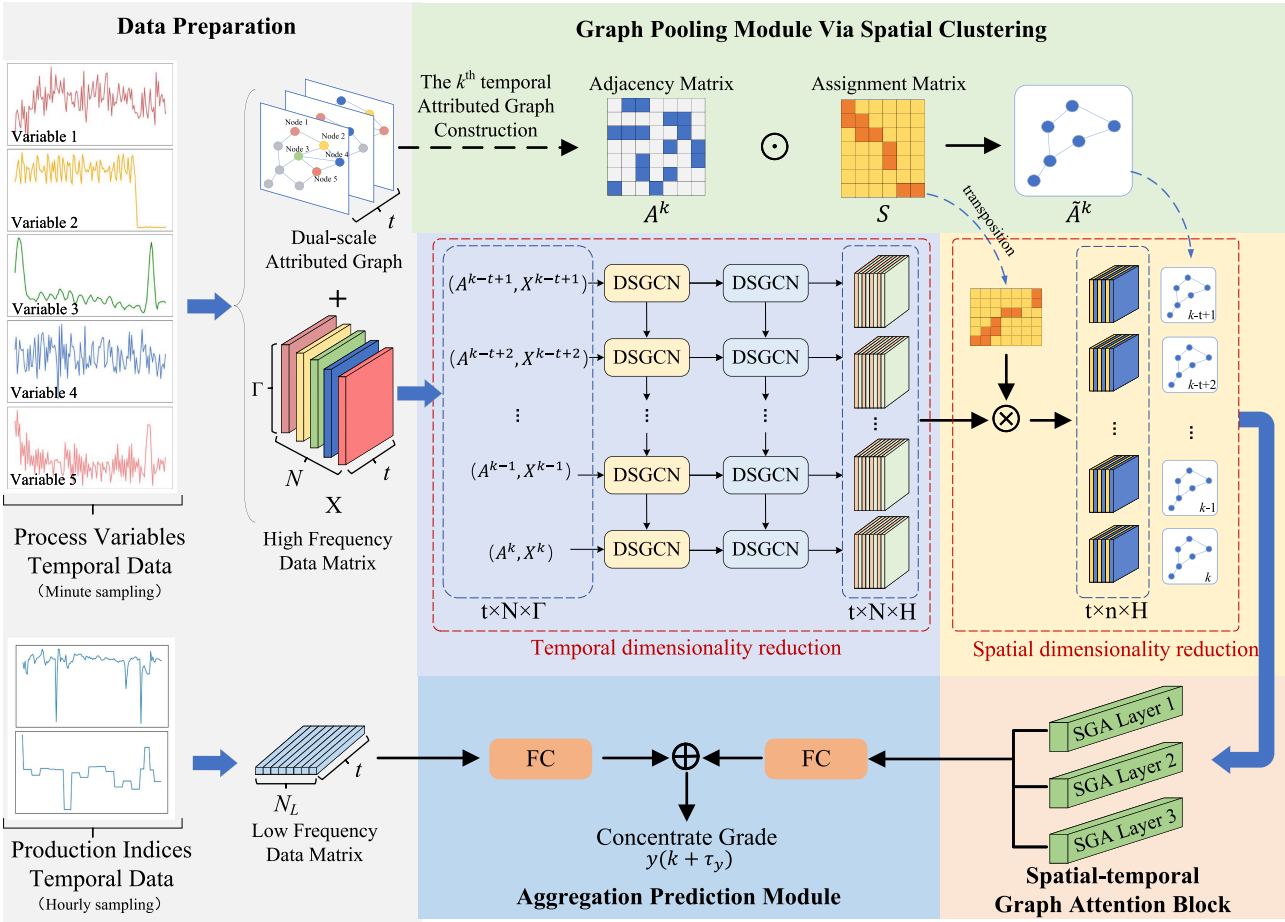

Fig. 4. The overall structure of DAGT for quality index prediction.

time. Industrial data is generated continuously and in real time, providing up-to-date information about the state and behavior of industrial processes. Industrial process data is often generated at high volumes and velocities, with large amounts of time series data being produced at frequent intervals.

- *Spatial Interactivity:* Industrial data is generated by sensors, instruments, and other devices distributed across different spatial locations within industrial production sites. The spatial interactivity of data involves the flow of data between different locations and the mutual influence of features. For instance, the order of input variables may determine the sequence in which information is processed by the model. Exploring the spatial arrangement characteristics of input variables can help models effectively capture spatial correlations in industrial data.

The dynamic characteristics that affect the quality index are hidden in the high and low-frequency industrial data. Due to the strong interconnections of the production process, establishing accurate quality index mathematical models is challenging. It is necessary to explore dual-scale spatial-temporal characteristics and model the nonlinear mapping relationship between the dual-scale industrial signal $[\boldsymbol{D}_H(k), \boldsymbol{D}_L(k)]$ and quality index predicted value $y(k+\tau_y)$.

III. Dual-Scale Attributed Graph Transformer

In this section, the dual-scale graph attribute structure for process variables data is defined with spatial-temporal characteristics. Then, an improved quality index prediction network is proposed, with its overall structure is shown in Fig. 4.

A. Dual-Scale Attributed Graph

In this paper, the dual-scale attributed graph is constructed first to describe the topology network and attribute features of process variables obtained by sampling at T_x interval in industrial processes.

Definition 1: data graph $\mathcal{G}^k$. Within the k^{th} production cycle, an undirected graph $\mathcal{G}^k = (A^k, X^k)$ is defined to describe the topology network of industrial process variables. The process variables $\Lambda = \{v_1, v_2, \ldots, v_N\}$ are regarded as nodes in the graph. The potential spatial relationships $A^k_{i,j}$ between node v_i and node v_j are regarded as edges in the data attributed graph, where $(v_i, v_j) \in \Lambda \times \Lambda$. The adjacency matrix $A^k \in \mathbb{R}^{N\times N}$ is used to represent the edge connections between all nodes on the graph topology network

Definition 2: feature matrix X^k. The process data on the graph topology network are regarded as the attribute feature of nodes, expressed as $X^k = [\boldsymbol{x}_1(k), \boldsymbol{x}_2(k), \ldots, \boldsymbol{x}_N(k)]^{\mathrm{T}} \in$

$\mathbb{R}^{N\times\Gamma}$. The dual-scale gap Γ is the graph features vector dimension, representing the number of node attribute features of process variables sampled by sampling period T_x within the k^{th} production cycle. $x_i = X^k[i,:] \in \mathbb{R}^{\Gamma}$ is the Γ-dimensional feature vector of node i, which represents the high-frequency sampling feature of the i^{th} process variable.

To define the graph structures between initial data, existing researchers usually use the connectivity between pairs of nodes to determine the adjacency matrix, such as the functional dependence between pairs of nodes [25], clustering and classification of nodes [26], cosine similarity between time series (e.g., Pearson Correlation Coefficients (PCCs) [27], MIC [22], [28]), and the shortest travel distance [29] between nodes if the initial data have geospatial properties.

Based on the characteristics of industrial data, the attributed graph is represented with a weighted directed adjacency matrix, which is constructed based on PCCs. The PCCs $\rho_{X,Y}$ is calculated as follows

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma X \sigma Y} \tag{5}$$

where $Cov(X,Y)$ is the covariance of X and Y, $\sigma X, \sigma Y$ are the standard deviation, $\rho_{X,Y} \in [-1.0, 1.0]$.

The lower the node correlation, the smaller the absolute value of $\rho_{X,Y}$. Assuming that when the positive and negative correlation $\rho_{X,Y}$ is less than the correlation factor ρ_a, there is no topology connection between nodes. It should be noted that in Section V, the optimal value of the correlation factor ρ_a is discussed and determined through ablation studies with different correlation factors. The adjacency matrix A^k is calculated as follows

$$A^k_{i,j} = \begin{cases} \rho_{X_i,Y_j}, & \text{if } |\rho_{X_i,Y_i}| \geq \rho_a, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $A^k_{i,j} = A^k[i,j] \in \mathbb{R}^{N\times N}$ is the adjacency relationship between node v_i and node v_j, $|\cdot|$ is the absolute operation.

Assuming a sliding time window is t, the dual-scale attributed graph data for the industrial process variables is denoted as $\mathcal{G} = [\mathcal{G}^{k-t+1}, \mathcal{G}^{k-t+2}, \ldots, \mathcal{G}^k]^{\mathrm{T}} \in \mathbb{R}^{t\times N\times\Gamma}$. The quality index prediction aims to predict the quality index value $y(k+\tau_y)$ for the $(\tau_y)^{\text{th}}$ step in the future, given attributed graph data of process variables and production indices data from the passed t time steps.

B. Dual-Scale Spatial-Temporal Graph Convolution Network (DSGCN)

GNNs learn feature representation for different nodes from attributed graph $\mathcal{G}^k = (A^k, X^k)$ using neighborhood aggregation schemes, which are formalized as the following passing function:

$$\hat{X}^k = GNN(X^k, A^k) \tag{7}$$

where X^k are the node features of the k^{th} temporal attributed graph, A^k is the adjacency matrix. k represents the k^{th} layer attributed graph.

Inspired by first-order graph Laplacian methods, the graph convolution operation of GCN is formally defined as:

$$\hat{X}^k = GCN(X^k, A^k) = ReLU(\tilde{D}^{-\frac{1}{2}}\hat{A}^k\tilde{D}^{-\frac{1}{2}}X^k\Theta^k) \tag{8}$$

where $\hat{A}^k = A^k + I$ is the adjacency matrix with self connections, I is the identity matrix. $\tilde{D}$ denotes the diagonal node degree matrix to normalize $\hat{A}^k$. $\hat{X}^k$ and X^k are the feature maps. Θ^k is a trainable matrix to perform feature transformation.

To consider the long and short-term temporal effects of attributed graphs, the slow-scale recurrent factors are used to construct a Dual-scale Spatial-temporal Graph Convolution Network (DSGCN), recording and transmitting the cell states of attributed graph features at fast-scale and slow-scale. This approach achieves temporal dimensionality reduction of high-frequency features while ensuring efficient feature extraction. The flow process of the temporal attributed graph X^k in the DSGCN is as follows:

$$\begin{aligned} X^{k'} &= LeakyReLU(GCN(X^k, A^k) \otimes s_{k-1}) \\ s_k &= \sigma(s_{k-1} \otimes X^k) \end{aligned} \tag{9}$$

where $X^{k'} \in \mathbb{R}^{N\times H}$ is the output sequence of unit, H is the hidden size of DSGCN, s_{k-1} is the recurrent factor state in the previous time step. $\otimes$ stands for element-wise product. $\sigma(\cdot)$ is the Sigmoid activation function. $LeakyReLU(\cdot)$ is a nonlinear activation function specifically designed to solve the dead-zone problem of ReLU, which is represented as follows

$$LeakyReLU(x) = \begin{cases} x, & x > 0, \\ ax, & x \leq 0. \end{cases} \tag{10}$$

where a is the negative slope (The default value of a is set to 0.2).

C. Graph Pooling Module via Spatial Clustering

Graph pooling methods are crucial for capturing the meaningful structure of an entire graph. Graph pooling aims to map the set of nodes into a compact representation, achieving spatial dimensionality reduction of data features. For example, the goal of graph pooling is to learn the relationships between $A^k \in \mathbb{R}^{N\times N}$, $X^{k'} \in \mathbb{R}^{N\times H}$ and $\tilde{A}^k \in \mathbb{R}^{n\times n}$, $\tilde{X}^k \in \mathbb{R}^{n\times H}$. The most straightforward graph pooling operation is to average all node characteristics. This method considers all node information equally without considering the graph's key features, which may lead to information loss. Therefore, several advanced pooling technologies have been proposed for the graph model recently, such as max-pooling [30], differentiable pooling [31], and structured pooling [32] based on cluster assignment, and have achieved good performance on multiple benchmark problems.

In the production process, the N nodes of the original graph $\mathcal{G}^k = (A^k, X^{k'})$ are distributed in n production cells $\{C_1, \ldots, C_n\}$. A clustering strategy based on process knowledge is used to group process variables in the adjacency matrix. Thus, the process variables can be divided into n clusters according to the location of the process space. Then, each cluster is transformed into a new node in the new graph $\tilde{\mathcal{G}}^k = (\tilde{A}^k, \tilde{X}^k)$. The cluster assignment can be an assignment matrix $S \in \mathbb{R}^{N\times n}$.

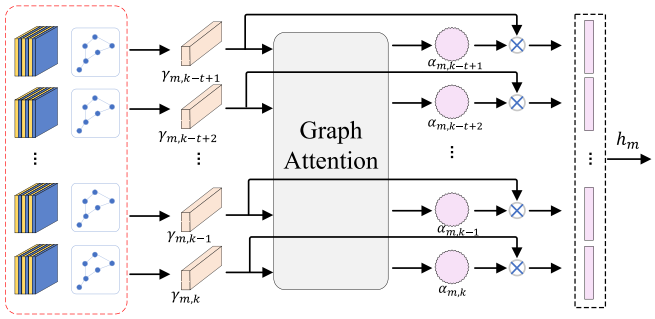

Fig. 5. Structure of the m^{th} spatial-temporal graph attention layer.

The assignment weight $s_{i,z} = S[i,z]$ can be formed as

$$s_{i,z} = \begin{cases} 1, & \text{if } v_i \text{ belongs to cluster } C_z, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

where $C_z \in \{C_1, \ldots, C_n\}$.

Then the new graph $\tilde{G}^k = (\tilde{A}^k, \tilde{X}^k)$ can be computed as

$$\begin{aligned} \tilde{X}^k &= S^{\mathrm{T}} X^{k'} \\ \tilde{A}^k &= S^{\mathrm{T}} A^k S \end{aligned} \tag{12}$$

D. Spatial-Temporal Graph Attention Block (SGAB)

The attention mechanism is usually applied to extract the key information from multiple sequence data. To extract feature information from the time series attributed graph, the Spatial-temporal Graph Attention Block (SGAB) is established based on attention mechanism [33] to perform spatial dimensionality reduction. To achieve sufficient expressive power, it is necessary to use a learnable linear transformation to convert the input compact features into higher-level features. At the m^{th} graph attention layer as shown in Fig. 5, the shared linear transformation weight matrix W_m^A, W_m^X is applied to each graph $(\tilde{A}^k, \tilde{X}^k)$ to construct graph sets $\gamma_{m,k}$.

$$\gamma_{m,k} = \tilde{X}^k W_m^X + \tilde{A}^k W_m^A + b_m \tag{13}$$

where $W_m^X \in \mathbb{R}^{H\times 1}, W_m^A \in \mathbb{R}^{n\times 1}$ and $b_m \in \mathbb{R}^{n\times 1}$ are shared trainable parameters at the m^{th} graph attention layer.

Then, an attention scoring function on the graph sets is performed to obtain attention scores $\alpha_{m,k}$. The LeakyReLU activation function, with negative slope $a = 0.2$, is applied to $\gamma_{m,k}$. The attention scores $\alpha_{m,k}$ is calculated by the *Softmax* activation function, as shown in (14).

$$\alpha_{m,k} = \frac{\exp(LeakyReLU\,(\gamma_{m,k}))}{\sum_{k=1}^{t} \exp(LeakyReLU\,(\gamma_{m,k}))} \tag{14}$$

The calculation process for the output of the m^{th} graph attention layer is as follows

$$h_m = \Big\Vert_{k=1}^{t} \alpha_{m,k} \otimes \gamma_{m,k} \tag{15}$$

where t is the time steps of input graph layers. $\|_{k=1}^{t}$ represents concatenation operation.

The multiple graph attention layers enhances the stability of the learning process within the self-attention mechanism. The average aggregation operations of multiple graph attention layers are expressed as

$$p_H = \frac{1}{M} \sum_{m=1}^{M} h_m \tag{16}$$

where M is the number of graph attention layers in SGAB.

IV. Analysis and Discussions

Graph theory is crucial in analyzing dual-scale time series industrial data for prediction tasks, particularly when considering the temporal continuity and spatial interactivity among variables. Graphs represent the temporal dependencies among industrial variables over time, with each variable as a node and edges denoting spatial correlations at different time points. An adjacency matrix is used to represent the edge structure. Each attributed graph includes low-frequency slow-scale data features, while each node within the graph contains high-frequency fast-scale data features.

The dual-scale attributed graph construction method based on PCCs aggregates information from neighboring nodes and learns representations that capture complex interactions and dependencies among industrial variables over time. By representing spatial-temporal dependencies, causal relationships, and complex interactions as graph structures, graph-based methods such as GCNs can extract spatial-temporal features, infer causal relationships, and enhance the accuracy of prediction models for quality index.

The process for predicting quality index using the DAGT method proposed in the paper can be broken down into three stages: temporal dimension reduction, spatial dimension reduction, and spatial-temporal graph feature extraction.

A. Temporal Dimensionality Reduction

GCNs are specialized neural networks designed to process graph-structured data. They utilize a message-passing scheme, where information flows from adjacent nodes to the central node under consideration. At each layer of the GCNs, messages from neighboring nodes are combined to create a new representation for the central node. allows GCNs to maintain spatial correlation features and fast-scale temporal characteristics within the attributed graph structure and high-frequency matrix. The DSGCN leverages GCNs to extract fast-scale spatial-temporal features while integrating slow-scale recurrent factors and nonlinear activation functions to transmit both long and short-term information features. This approach achieves temporal dimensionality reduction of high-frequency features while ensuring efficient feature extraction.

B. Spatial Dimensionality Reduction

The graph pooling module employs spatial clustering to transform the adjacency matrix and input features. A clustering strategy based on process knowledge is used to group variables in the adjacency matrix, enabling a more efficient representation

of the graph's spatial structure. This approach groups adjacent nodes within the same subprocess in the industrial process space to create an assignment matrix. By aggregating the assignment matrix with the attributed graph, spatial dimensionality reduction of high-dimensional process variables is achieved. The graph pooling module reduces the computational complexity of the model while preserving key spatial information.

C. Spatial-Temporal Graph Features Extraction

After reducing temporal and spatial dimensions, a shared linear transformation matrix aggregates attributed graph $\tilde{G}^k = (\tilde{A}^k, \tilde{X}^k)$ information into a graph set $\gamma_{m,k}$. The attention scoring function then selectively focuses on key areas of the attributed graph sequence, assigning greater weight to regions with important spatial-temporal features. The attention scoring function employs the *Softmax* function to normalize attention scores across sequence elements, ensuring that the scores sum to one and form a probability distribution over the sequence elements. This probabilistic interpretation facilitates an intuitive understanding of the attention mechanism.

The proposed DAGT fully leverages the characteristics of inconsistent sampling frequencies in industrial data by using the dual-scale gap of high and low-frequency sampling as the feature dimension of the attributed graph for high-frequency process variables, thereby defining a new dual-scale attributed graph. As described above, DAGT achieves dimensionality reduction in the temporal and spatial dimensions of high-frequency process data through DSGCN, graph pooling module via spatial clustering, and SGAB, thereby obtaining spatial-temporal graph features.

In the aggregation prediction module, fully connected layers convert slow-scale input data and spatial-temporal graph features of fast-scale data into higher-level representations. These layers learn complex combinations of features from original representations, maintaining a strong correlation between the dual-scale industrial data and the predicted quality index. These strategies enable DAGT to integrate and utilize information from both fast-scale and slow-scale data, leading to more accurate and comprehensive modeling of industrial processes.

V. Industrial Application

A. Mineral Processing Process

After the raw ore is extracted from underground, it undergoes physical and chemical reactions such as crushing and magnetic separation to enrich useful minerals, in order to obtain qualified concentrate and tailings [4], [34]. The whole production process of mineral processing consists of five sub-processes (SP1-SP5): (SP1) crushing and screening, (SP2) high-pressure roll grinding, (SP3) grinding and magnetic separation, (SP4) tailings settlement, and (SP5) concentrate dewatering. As illustrated in Fig. 6, the raw ore is roughly crushed by jaw crusher and cone crusher to obtain waste ore and crushed ore in SP1. The crushed ore is rolled by a high-pressure roller mill and screened by vibrating machines in SP2 to produce rolled ore with a particle size less than 3mm. Then, these rolled ores are mixed with water

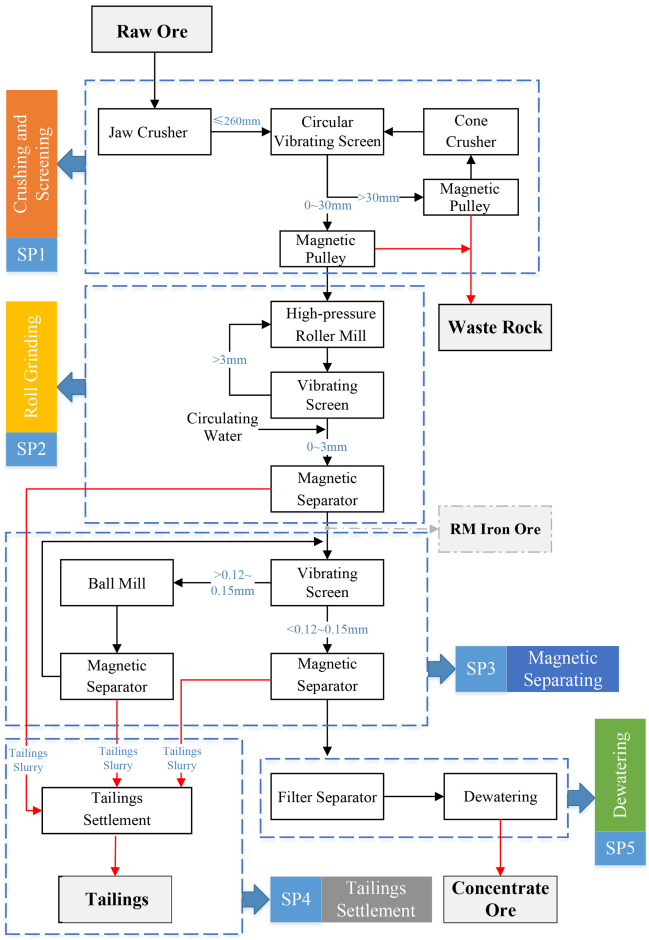

Fig. 6. Flow chart of mineral processing process.

and ground again by a ball mill, and separated by magnetic separators to obtain high-grade concentrate. The tailings are mixed, concentrated and settled in SP4. Finally, the concentrate ore with the desired particle size is concentrated and dehydrated in SP5 to produce the final concentrate product. Concentrate grade is the final quality index in the mineral processing process.

B. Data Processing

In our case, the experimental data is the industrial data of a beneficiation plant for the past two years. The annual processing capacity of the beneficiation plant exceeds 3 million tons. High-frequency minute sampling data ($T_x = 1$ minute) is acquired by DAC, and low-frequency hourly sampling data ($T_p = 60$ minutes) is obtained by the Data-Recording System (DRS). The dual-scale gap is $\Gamma = T_p/T_x = 60$. After removing the unproduced empty data, there are 595320 sets of high-frequency production data, and the number of high-frequency process variables is $N = 18$. There are 9922 sets of low-frequency production data, and the number of low-frequency production indices is $N_L = 11$. This paper implements time matching for high and low-frequency data based on hourly sampling timestamps. Each group of sequences is composed of high-frequency data $D_{H,t} \in \mathbb{R}^{1\times(N\times\Gamma)}$ and low-frequency data $D_{L,t} \in \mathbb{R}^{1\times N_L}$. The sequence length is 9922, with a data split of 8012 data groups

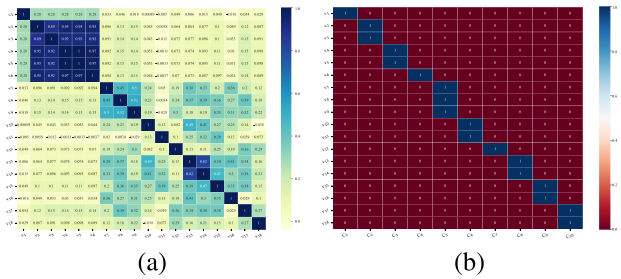

Fig. 7. Correlation coefficient and assignment matrix. (a) is the PCCs of process variables and (b) is the assignment matrix S obtained by spatial clustering based on the process knowledge.

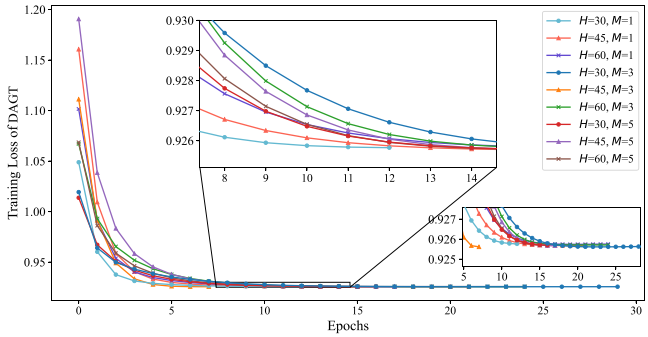

Fig. 8. Training loss of DAGT with different hyper-parameters combinations.

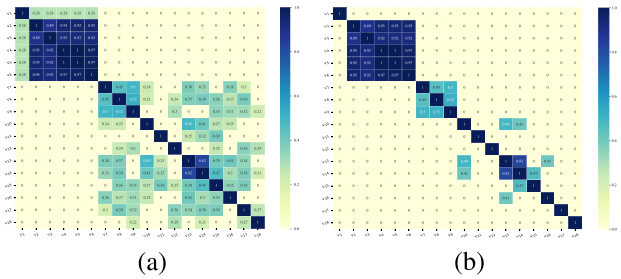

Fig. 9. Correlation coefficient and structural matrix. (a) is the adjacency matrix A^k with $\rho_a = 0.2$ and (b) is the adjacency matrix A^k with $\rho_a = 0.4$.

designated for the training set, 1415 data points allocated to the validation set, and 495 data points reserved for the test set.

The PCCs of process variables in the k^{th} production cycle calculated by (5) is shown in Fig. 7(a). According to the distribution of the production closed-loop control system, the five sub-processes (SP1-SP5) of the mineral processing process are divided into ten production units. The process variables $\Lambda = \{v_1, v_2, \ldots, v_N\}$ are divided into ten production units ($n = 10$) according to the different sampling positions. The assignment matrix $S \in \mathbb{R}^{18\times10}$ obtained by spatial clustering based on process knowledge is shown as Fig. 7(b).

C. Quality Index Prediction

The full structure of the Dual-scale Attributed Graph Transformer, comprising DSGCN, graph pooling module via spatial clustering, and SGAB, is now described. The Encoder: $\mathcal{G}^k = (A^k, X^k) \to \hat{X}^k \in \mathbb{R}^{N\times H}$ is denoted as follows:

$$Encoder(X^k, A^k) = DSGCN_2(DSGCN_1(X^k, A^k)) \quad (17)$$

where two layers of DSGCN are stacked to capture spatial-temporal features. The output in k^{th} time step of the Encoder is $\hat{X}^k$.

After that, the graph pooling module and SGAB aggregate the features into a single matrix p_H.

$$p_H = SGAB(Pooling(\hat{\boldsymbol{X}}, \boldsymbol{A})) \quad (18)$$

where $\hat{\boldsymbol{X}} = [\hat{X}^{k-t+1}; \ldots; \hat{X}^k]$, $\boldsymbol{A} = [A^{k-t+1}; \ldots; A^k]$.

Two fully connected (FC) layers in the aggregation prediction module are utilized to combine the low-frequency production indices data $p_L \in \mathbb{R}^{t\times N_L}$ and the attributed graph feature p_H to produce the predicted quality index $y(k+\tau_y)$.

$$y(k+\tau_y) = W_L p_L + b_L + W_H p_H + b_H \quad (19)$$

where W_L is the weight matrix connected to the low-frequency vector p_L. W_H is the weight matrix connected to the p_H. b_L, b_H are the bias parameters of the FC layers.

In this paper, root mean square error (RMSE) and mean absolute error (MAE) are used as the evaluation metrics of the concentrate grade prediction model. They are defined as follows:

$$RMSE = \sqrt{\frac{1}{l_{test}}\sum_{i=1}^{l_{test}}(Y_i - \hat{Y}_i)^2} \quad (20)$$

$$MAE = \frac{1}{l_{test}}\sum_{i=1}^{l_{test}}\left|Y_i - \hat{Y}_i\right| \quad (21)$$

where l_{test} is the length of test data, Y_i is the i^{th} true value of the test data, $\hat{Y}_i$ represents the i^{th} predicted value of the test data.

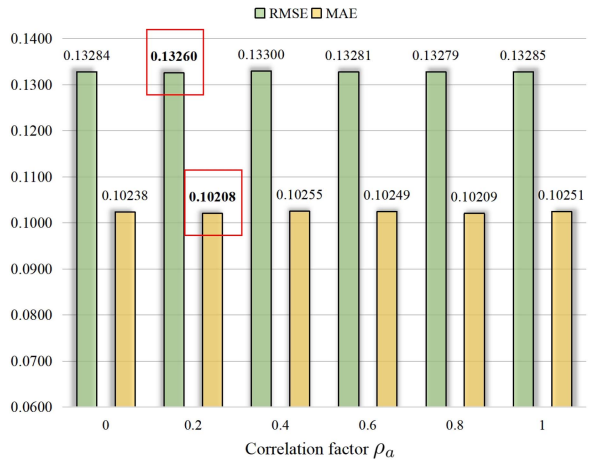

Fig. 10. RMSE and MAE results of DAGT with different correlation factors.

D. Hyper-Parameters Analysis

In the hyper-parameters grid search experiment of DAGT, the hidden size H of DSGCN is selected from the set {30, 45, 60}, and the number M of layers in SGAB is selected from the set {1, 3, 5}. The training loss curves with different hyper-parameters

TABLE I
HYPER-PARAMETERS AND PARAMETER CONFIGURATION FOR DAGT MODEL

Parameters	Value	Annotation
H	30	Hidden size of TGCN
M	3	Number of graph attention layers
t	5	Sliding time steps
n	10	Number of production cells
ρ_a	0.2	Correction factor
N	18	Number of nodes (process variables)
N_L	11	Number of low-frequency indices
Γ	60	Dual-scale gap

combinations are illustrated in Fig. 8. $H = 30, M = 3$ in DAGT is the optimal hyper-parameters combination and has the lowest RMSE of prediction performance. It is observed that the optimal hyper-parameters combination has the maximum number of training iterations and significantly reduces the risk of model overfitting. The model training for other hyper-parameter combinations is prematurely terminated due to a lack of decrease in the validation loss. Furthermore, the initial iteration of the model training loss is consistently lower when $H = 30$ compared to other H. In the DAGT, selecting hyper-parameters is essential to improve the training efficiency and prediction accuracy. The choice of H may be affected by the dual-scale gap Γ. The relationship between the hyper-parameters and the dual-scale gap will be investigated in future research.

This work considered two nodes with a correlation coefficient less than ρ_a as having no topological connection when building DAGT. For example, Fig. 9(a) and (b) show the adjacency matrix A^k of the attributed graph with $\rho_a = 0.2$ and $\rho_a = 0.4$, respectively. It can be seen that different correlation factors determine the complexity of the topology between nodes. The smaller the ρ_a, the more complex the topology relationship of nodes.

To deeply explore the importance of different topological relationships between nodes for model accuracy, ablation studies with different correlation factors are performed to determine the optimal value of the correlation factor ρ_a. The RMSE and MAE results of ablation studies are shown in Fig. 10. The experimental results show that when $\rho_a = 0.2$, the modeling accuracy of DAGT is better, and it can further explore the dual-scale characteristics. The hyper-parameters grid search result and related parameters configuration is shown in Table I.

E. Comparison

The proposed DAGT method is compared with eight typical time series modeling methods in this comparative experiment, including the GRU network [7], Bidirectional GRU network, LSTM network [5], Bidirectional LSTM network, ConvLSTM network [35], SIDL network [36], LSTM-ES [37] and HFLF-LSTM [3]. The comparison methods use the full connection layer as the output layer. In the ConvLSTM network, the convolutional kernel has a size of 3×3 and a stride of 1. In the SIDL network, the low-frequency production indices are used as the input of identifiable linear model. The unknown high-order nonlinear model is trained using a stacked LSTM network with all industrial data. The sum of linear identification model and nonlinear model is the final output of SIDL. In the LSTM network with expansion and serialization of time series data (LSTM-ES), it extracts high-frequency data from the last 20 minutes of each hour, expanding and serializing them to match the time series of low-frequency data. These new data objects are then inputted into a stacked LSTM network for prediction. In the HFLF-LSTM, it constructs high frequency unit and low frequency unit to extract dual-scale industrial data features, and the new features are fed into a stacked LSTM network for prediction. To standardize the deep network structure or stacked LSTM structure of the comparison methods, the number of neurons is fixed at 5. The hidden size in a single neuron is set to 30. The number of layers is set to 3.

TABLE II
PERFORMANCE EVALUATION OF CONCENTRATE GRADE **one-step**($\tau_y = 1$) PREDICTION EXPERIMENT WITH DIFFERENT METHODS

Methods	RMSE	MAE	$[-0.3\%, 0.3\%]$	$[-0.1\%, 0.1\%]$
GRU	0.1594	0.1262	80.40%	35.15%
BiGRU	0.1485	0.1193	84.44%	35.76%
LSTM	0.1570	0.1222	82.42%	35.56%
BiLSTM	0.1568	0.1234	81.41%	33.93%
ConvLSTM	0.1557	0.1218	84.04%	35.35%
SIDL	0.1524	0.1192	84.04%	35.76%
LSTM-ES	0.2150	0.1692	69.09%	25.05%
HFLF-LSTM	0.1443	0.1128	84.04%	39.60%
DAGT	**0.1326**	**0.1022**	**88.89%**	**46.06%**

In the comparative experiment, the Adam optimizer is employed with a learning rate of 0.001. The batch size for all methods is set to 64. The maximum number of iterations is set to 30. Training stops to prevent overfitting when the validation loss does not decrease for five consecutive training stages.

F. Results and Discussions

In the industrial experiment of concentrate grade one-step ($\tau_y = 1$) prediction, the prediction results of the proposed method and eight comparative methods are shown in Fig. 11(a)–(h). It can be seen that the fitting trend of the prediction method proposed in this paper has a better prediction effect than the comparison methods. The comparison details of prediction curves for concentrate grade based ConvLSTM, SIDL, LSTM-ES, HFLF-LSTM and DAGT are shown in Fig. 12. The proposed DAGT not only has a better prediction trend but also can avoid the decline of prediction accuracy caused by the fluctuation of industrial data.

To further quantify the prediction performance of different methods, this work calculates the model evaluation metrics and the interval qualification rate of all methods, listed in Table II. The interval qualification rate refers to the deviation ratio between the predicted value and the true value within the specified absolute deviation interval (e.g. 0.3% and 0.1%). The comparison results in Table II show that the proposed DAGT method predicts the concentrate grade well, the prediction evaluation metrics of which are superior to the comparison methods.

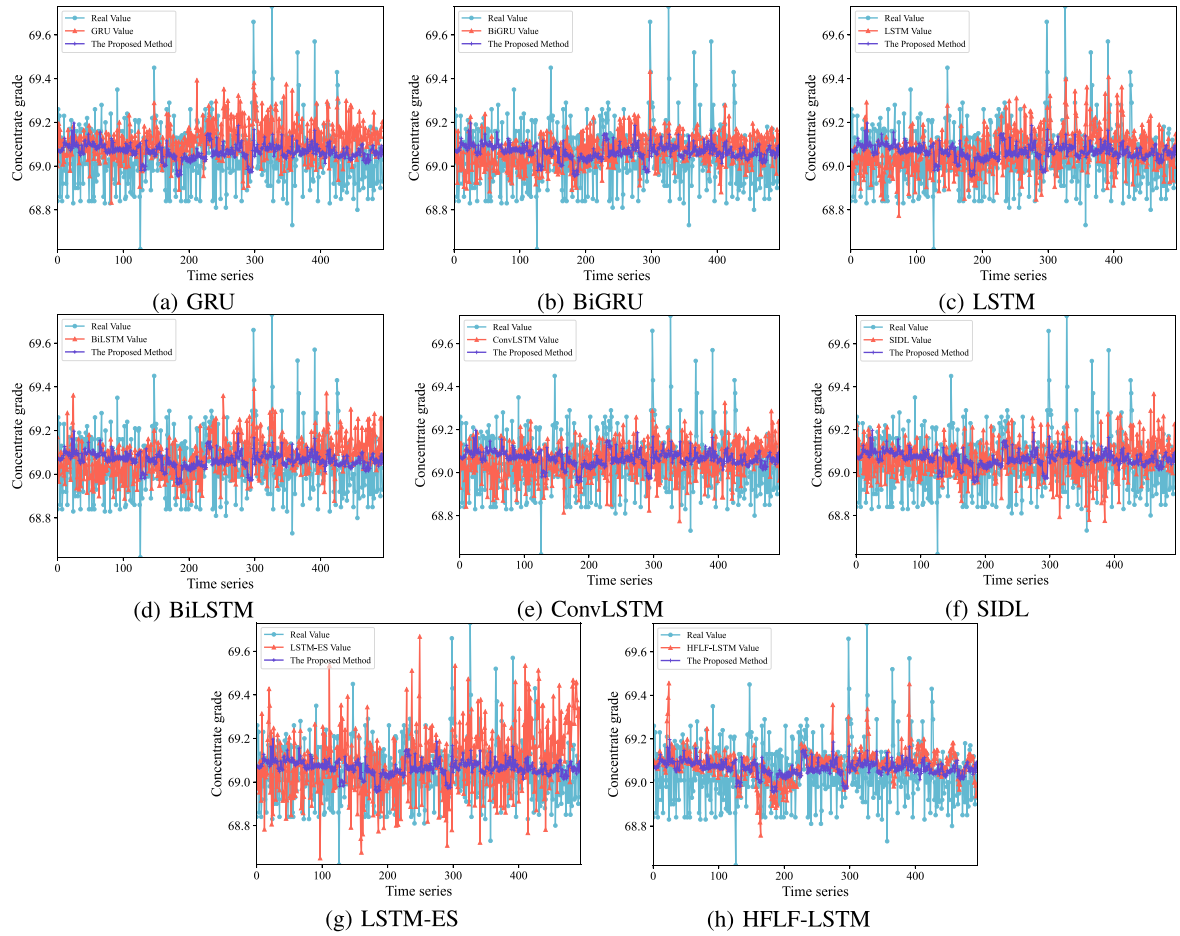

Fig. 11. Prediction curves of different comparison methods and the proposed DAGT.

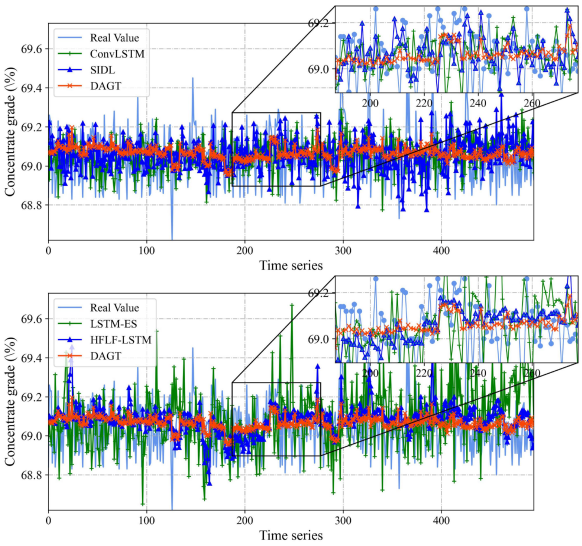

Fig. 12. Comparison details of prediction curves for concentrate grade based on different methods and DAGT.

In the test data with a length of 495, there are 440 samples at $[-0.3\%, 0.3\%]$ and 228 samples at $[-0.1\%, 0.1\%]$ in the DAGT prediction experiment. The concentrate grade one-step prediction error of DAGT is plotted in Fig. 13. The proposed method's 0.3% prediction interval qualification rate is 88.89%, and the proposed method's 0.1% prediction interval qualification rate is 46.06%.

TABLE III
PERFORMANCE EVALUATION OF CONCENTRATE GRADE **multi-step**($\tau_y = 5$) PREDICTION EXPERIMENT WITH DIFFERENT METHODS

Methods	RMSE	MAE	$[-0.3\%, 0.3\%]$	$[-0.1\%, 0.1\%]$
GRU	0.1539	0.1205	85.25%	39.60%
BiGRU	0.1511	0.1184	83.43%	37.98%
LSTM	0.1454	0.1138	87.47%	36.77%
BiLSTM	0.1537	0.1174	85.25%	38.99%
ConvLSTM	0.1540	0.1183	86.87%	40.00%
SIDL	0.1559	0.1219	83.23%	39.19%
LSTM-ES	0.2045	0.1582	74.34%	30.51%
HFLF-LSTM	0.1469	0.1117	85.66%	39.19%
DAGT	**0.1333**	**0.1015**	**88.48%**	**48.08%**

Similarly, Table III shows evaluation results of concentrate grade multi-step ($\tau_y = 5$) prediction experiment with different methods. The proposed method has the best prediction performance compared with other methods. The proposed method's 0.3% prediction interval qualification rate is 88.48%, and its 0.1% prediction interval qualification rate is 48.08%.

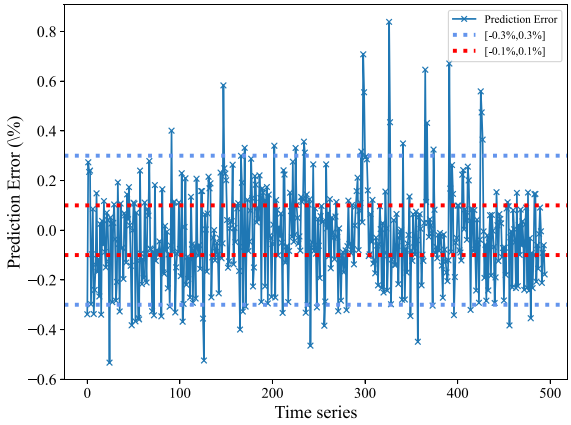

Fig. 13. The concentrate grade one-step prediction error of DAGT.

VI. Conclusion

This paper addressed the challenge of inaccurate prediction due to inconsistent data sampling frequencies in complex industrial processes. To overcome this limitation, this work proposed a novel deep learning architecture, the Dual-scale Attribute Graph Transformer (DAGT), for effectively extracting spatial-temporal features from attributed graph data. DAGT addresses inconsistent sampling frequencies in industrial data streams by introducing the Dual-Scale Spatial-temporal Graph Convolution Network (DSGCN) and the Spatial-temporal Graph Attention Block (SGAB). DSGCN uses both spatial and temporal information within attributed graphs, while SGAB employs an attention mechanism to prioritize crucial regions of the graph sequence. The effectiveness of DAGT is rooted in its unique design elements, including the dual-scale adjacency matrix for efficient dimensionality reduction and the graph pooling module via spatial clustering. These innovations enable DAGT to learn robust feature representations from attributed graph sequences.

The experiments conducted in this paper demonstrate the efficacy of DAGT. However, the significance of this work extends beyond the immediate application. The proposed framework offers a general-purpose solution for tasks requiring spatial-temporal feature extraction from attributed graphs. Future research directions include exploring DAGT's applicability in various domains beyond industrial process data analysis and potentially investigating extensions to handle even more complex graph structures.

References

[1] T. Chai, J. Ding, G. Yu, and H. Wang, "Integrated optimization for the automation systems of mineral processing," *IEEE Trans. Automat. Sci. Eng.*, vol. 11, no. 4, pp. 965–982, Oct. 2014.

[2] J. Ding, H. Modares, T. Chai, and F. L. Lewis, "Data-based multiobjective plant-wide performance optimization of industrial processes under dynamic environments," *IEEE Trans. Ind. Inform.*, vol. 12, no. 2, pp. 454–465, Apr. 2016.

[3] K. Zhang, Y. Yu, Y. Jia, and T. Chai, "Comprehensive production index prediction using dual-scale deep learning in mineral processing," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 23, 2024, doi: 10.1109/TNNLS.2024.3421570.

[4] J. Ding, C. Yang, and T. Chai, "Recent progress on data-based optimization for mineral processing plants," *Engineering*, vol. 3, no. 2, pp. 183–187, 2017, doi: 10.1016/J.ENG.2017.02.015.

[5] T. Wang, H. Leung, J. Zhao, and W. Wang, "Multiseries featural LSTM for partial periodic time-series prediction: A case study for steel industry," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 5994–6003, Sep., 2020.

[6] P. Ma, G. Li, H. Zhang, C. Wang, and X. Li, "Prediction of remaining useful life of rolling bearings based on multiscale efficient channel attention CNN and bidirectional GRU," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 2508413.

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[8] J. Sun, X. Meng, and J. Qiao, "Prediction of oxygen content using weighted PCA and improved LSTM network in MSWI process," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 2507512.

[9] J. Geng, C. Yang, Y. Li, L. Lan, and Q. Luo, "MPA- RNN: A. novel attention-based recurrent neural networks for total nitrogen prediction," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 6516–6525, Oct. 2022.

[10] A. N. Jahromi, S. Hashemi, A. Dehghantanha, R. M. Parizi, and K. -K. R. Choo, "An enhanced stacked LSTM method with no random initialization for malware threat hunting in safety and time-critical systems," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 4, no. 5, pp. 630–640, Oct. 2020.

[11] M. M. Rahman and Y. Watanobe, "Multilingual program code classification using n-Layered Bi-LSTM model with optimized hyperparameters," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 8, no. 2, pp. 1452–1468, Apr. 2024.

[12] F. Zhou, Q. Yang, T. Zhong, D. Chen, and N. Zhang, "Variational graph neural networks for road traffic prediction in intelligent transportation systems," *IEEE Trans. Ind. Inform.*, vol. 17, no. 4, pp. 2802–2812, Apr. 2021.

[13] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[14] M. Jin et al., "A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection," 2023. Accessed: Mar. 2024.

[15] X. Li, Y. Jiang, Y. Liu, J. Zhang, S. Yin, and H. Luo, "RAGCN: Region aggregation graph convolutional network for bone age assessment from X-ray images," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 4006412.

[16] T.N. Kipf and M. Welling, " Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017.

[17] 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. 01, 2020, pp. 914–921, doi: 10.1609/aaai.v34i01.54380.

[18] Y. Ma, H. Lou, M. Yan, F. Sun, and G. Li, "Spatio-temporal fusion graph convolutional network for traffic flow forecasting," *Inf. Fusion*, vol. 104, 2024, doi: 10.1016/j.inffus.2023.102196.

[19] Y. Xiao et al., "AFSTGCN: Prediction for multivariate time series using an adaptive fused spatial-temporal graph convolutional network," *Digit. Commun. Netw.*, vol. 10, no. 2, pp. 292–303, Apr. 2024.

[20] B. Yu, H. Xie, M. Cai, and W. Ding, " MG-GCN: Multi-granularity graph convolutional neural network for multi-label classification in multi-label information system," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 8, no. 1, pp. 288–299, Feb. 2024.

[21] J. Li, Y. Shi, H. Li, and B. Yang, "TC-GATN: Temporal causal graph attention networks with nonlinear paradigm for multivariate time-series forecasting in industrial processes," *IEEE Trans. Ind. Inform.*, vol. 19, no. 6, pp. 7592–7601, Jun. 2023, doi: 10.1109/TII.2022.3211330.

[22] Y. Song, D. Tang, J. Yu, Z. Yu, and X. Li, "Short-term forecasting based on graph convolution networks and multiresolution convolution neural networks for wind power," *IEEE Trans. Ind. Inform.*, vol. 19, no. 2, pp. 1691–1702, Feb. 2023.

[23] H. Wang, R. Liu, S. X. Ding, Q. Hu, Z. Li, and H. Zhou, "Causal-trivial attention graph neural network for fault diagnosis of complex industrial processes," *IEEE Trans. Ind. Inform.*, vol. 20, no. 2, pp. 1987–1996, Feb. 2024.

[24] K. Zhang, Q. Xu, C. Liu, and T. Chai, "Intelligent decision-making system for mineral processing production indices based on digital twin interactive visualization," *J. Visual.*, vol. 27, pp. 421–436, 2024, doi: 10.1007/s12650-024-00964-4.

[25] Y. Wang, Z. Duan, Y. Huang, H. Xu, J. Feng, and A. Ren, "MTHetGNN: A heterogeneous graph embedding framework for multivariate time series forecasting," *Pattern Recognit. Lett.*, vol. 153, pp. 151–158, 2022.

[26] P. Lang, X. Fu, J. Dong, H. Yang, and J. Yang, "A novel radar signals sorting method via residual graph convolutional network," *IEEE Signal Process. Lett.*, vol. 30, pp. 753–757, 2023.
[27] P. Khosravinia, T. Perumal, and J. Zarrin, "Enhancing road safety through accurate detection of hazardous driving behaviors with graph convolutional recurrent networks," *IEEE Access*, vol. 11, pp. 52983–52995, 2023, doi: 10.1109/ACCESS.2023.3280473.
[28] T. Zhang, C. Liu, Z. Liu, J. Tan, and M. Ahmat, "Temporal double graph convolutional network for CO and CO prediction in blast furnace gas," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 2502113.
[29] X. Zhang, R. Cao, Z. Zhang, and Y. Xia, "Crowd flow forecasting with multi-graph neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, Glasgow, U.K., 2020, pp. 1–7, doi: 10.1109/IJCNN48605.2020.9207457.
[30] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2013, *arXiv:1312.6203*.
[31] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4800–4810.
[32] H. Yuan and S. Ji, "StructPool: Structured graph pooling via conditional random fields," in *Proc. Int. Conf. Learn. Representations*, 2020.
[33] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.
[34] Q. Xu, K. Zhang, M. Li, Y. Chu, and D. Zhang, "Multi-objective robust optimization for planning of mineral processing under uncertainty," in *Proc. 33rd Chin. Control Decis. Conf.*, Kunming, China, 2021, pp. 4020–4027.
[35] H. Huang, Z. Zeng, D. Yao, X. Pei, and Y. Zhang, "Spatial-temporal ConvLSTM for vehicle driving intention prediction," *Tsinghua Sci. Technol.*, vol. 27, no. 3, pp. 599–609, Jun. 2022, doi: 10.26599/TST.2020.9010061.
[36] T. Chai, J. Zhang, and T. Yang, "Demand forecasting of the fused magnesia smelting process with system identification and deep learning," *IEEE Trans. Ind. Inform.*, vol. 17, no. 12, pp. 8387–8396, Dec. 2021.
[37] Y. Li, C. Yang, and Y. Sun, "Dynamic time features expanding and extracting method for prediction model of sintering process quality index," *IEEE Trans. Ind. Inform.*, vol. 18, no. 3, pp. 1737–1745, Mar. 2022.

Kesheng Zhang received the B.E. degree in automation in 2018 from Northeastern University, Shenyang, China, where he is currently working toward the Ph.D. degree in control science and engineering with the State Key Laboratory of Synthetical Automation for Process Industries. His research interests include industrial modeling and optimization, deep learning, deep reinforcement learning, intelligent systems, and visual analysis for complex industrial processes.

Wen Yu (Senior Member, IEEE) received the B.S. degree in automation from Tsinghua University, Beijing, China, in 1990, and the M.S. and Ph.D. degrees in automatic control from Northeastern University, Shenyang, China, in 1992 and 1995, respectively. From 1995 to 1996, he was a Lecturer with the Department of Automatic Control, Northeastern University. Since 1996, he has been with CINVESTAV-IPN (National Polytechnic Institute), Mexico City, Mexico, where he is currently a Professor with the Departamento de Control Automatico. From 2002 to 2003, he held research positions with the Instituto Mexicano del Petroleo, Mexico City. He was a Senior Visiting Research Fellow with Queen's University Belfast, Belfast, U.K., from 2006 to 2007, and a Visiting Associate Professor with the University of California, Santa Cruz, CA, USA, from 2009 to 2010. He has been a Visiting Professorship with Northeastern University since 2006. He holds the distinguished position of a Full Professor (Investigador Cinvestav 3F) with the Department de Automatic Control, CINVESTAV-IPN. He is a Member of the Mexican Academy of Sciences, Mexico. He has more than 500 publications, including more than 200 journal papers and eight monographic books. He has supervised 37 Ph.D. theses and 38 master theses. Several of which have received awards in national competitions. His publications currently report 10,714 citations and his h-index is 52 according to Google Scholar. He is among the top 2% of the most-cited scientists in the world Stanford/Elsevier, 2023. On Research.com's list of World's Best Scientists, in electronics and electrical engineering as well as computer science, he holds the 6th position and 5th position in Mexico. He was the General Chair of the IEEE flagship annual meeting SSCI 2023. He is an Associate Editors for IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Neurocomputing*, and *Journal of Intelligent and Fuzzy Systems*.

Tianyou Chai (Life Fellow, IEEE) received the Ph.D. degree in control theory and engineering from Northeastern University, Shenyang, China, in 1985. He became a Professor with Northeastern University, in 1988. He is the Founder and Director of the Center of Automation, which became a National Engineering and Technology Research Center and a State Key Laboratory. He has authored or coauthored 297 peer-reviewed journal articles. His research interests include modeling, control, optimization, and integrated automation of complex industrial processes. Dr. Chai is a Member of the Chinese Academy of Engineering and an IFAC Fellow. His paper titled Hybrid Intelligent Control for Optimal Operation of Shaft Furnace Roasting Process was selected as one of the three best papers for the Control Engineering Practice Paper Prize for 2011Â¨C2013. He has developed control technologies with applications to industrial processes. For his contributions, he has won five prestigious awards of National Natural Science, National Science and Technology Progress and National Technological Innovation, 2007 Industry Award for Excellence in Transitional Control Research from IEEE Multiple-Conference on Systems and Control, and 2017 Wook Hyun Kwon Education Award from Asian Control Association. He was the Director of the Department of Information Science, National Natural Science Foundation of China, from 2010 to 2018.