SCAT: A Time Series Forecasting with Spectral Central Alternating Transformers

Chengjie Zhou[1] , **Chao Che**[2] , **Pengfei Wang**[1] and **Qiang Zhang**[1*]

[1]School of Computer Science and Technology, Dalian University of Technology
[2]Key Laboratory of Advanced Design and Intelligent Computing (Dalian University), Ministry of Education, Dalian University
millzhou@mail.dlut.edu.cn, chechao@gmail.com, wangpf@dlut.edu.cn, zhangq@dlut.edu.cn

Abstract

Time series forecasting has essential applications across various domains. For instance, forecasting power time series can optimize energy usage and bolster grid stability and reliability. Existing models based on transformer architecture are limited to classical design, ignoring the impact of spatial information and noise on model architecture design. Therefore, we propose an atypical design of Transformer-based models for multivariate time series forecasting. This design consists of two critical components: (i) spectral clustering center of time series employed as the focal point for attention computation; (ii) alternating attention mechanism wherein each query transformer is compatible with spectral clustering centers, executing attention at the sequence level instead of the token level. The alternating design has a two-fold benefit: firstly, it eliminates the uncertainty noise present in the dependent variable sequence of the channel input, and secondly, it incorporates the Euclidean distance to mitigate the impact of extreme values on the attention matrix, thereby aligning predictions more closely to the sequence's natural progression. Experiments on ten real-world datasets, encompassing Wind, Electricity, Weather, and others, demonstrate that our Spectral Central Alternating Transformer (SCAT) outperforms state-of-the-art methods (SOTA) by an average of 17.5% in power time series forecasting.

1 Introduction

Time-series forecasting[Box *et al.*, 2015] is one of the most important renewable energy power-generation technologies. Moreover, carbon neutrality[Khalifa *et al.*, 2022] is a global goal for realizing the sustainable development of the world's resources. However, wind[Moradzadeh *et al.*, 2021] and solar power[Doubleday *et al.*, 2020] stations need more accurate power prediction algorithms[Wang *et al.*, 2011; Elsaraiti and Merabet, 2022; Voyant *et al.*, 2017] to improve energy utilization, which can support the rapid growth of power-generation capacity. As depicted in Figure 1, wind and solar power generation can be forecasted using input features such as weather conditions.

With the rapid proliferation of artificial intelligence, research focused on time-series forecasting[Le Guen and Thome, 2020] utilizing deep learning models has surged. In this array of models, the transformer architecture, initially devised for natural language processing (NLP), has emerged as a versatile tool with applications extending to diverse domains, including audio processing[Chan *et al.*, 2015], computer vision (CV)[Dosovitskiy *et al.*, 2020], and time-series analysis[Lim and Zohren, 2021]. Transformers have exhibited remarkable capabilities in capturing dependencies among elements and unveiling latent relationships among tokens within sequences. More importantly, various transformer-based variants, such as Autoformer, Reformer, Pyraformer, and ETSformer, have proved their mettle in time-series forecasting. These models have pushed the boundaries of prediction accuracy by leveraging innovative techniques like series decomposition.

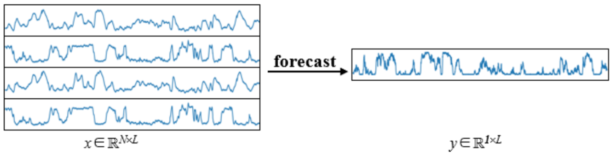

Figure 1: The power time series forecasting process entails inputting a group of variables, including wind speed and direction. SCAT is applied to produce an output signal for future power series.

This paper introduces a novel concept involving utilizing spectral clustering centers as pivotal components in attention mechanisms for time-series data. The proposed spectral central alternating transformer (SCAT) encompasses two main innovative designs:

- **Series Clustering Center.** Time-series data often comprise diverse components, with examples like wind and solar generation data encompassing weather-related information and generation capacity metrics. However, weather-related data often exhibit inaccuracies stemming from limitations in data acquisition equipment or challenges in advanced weather prediction techniques.

[*]Corresponding author.

Notably, previous research has paid limited attention to the adverse impact of noise present in these sequences, particularly at the local level, which can significantly affect prediction accuracy. In contrast, SCAT addresses this issue by employing a preprocessing step that involves clustering and classifying the meteorological time-series data. This process identifies and separates central points within each sub-sequence, each representing distinct weather patterns. This crucial preprocessing step occurs before the primary model architecture is computed. Subsequently, we utilize these spectral clustering centers as computational cores to eliminate the redundancy of traditional attention computing structures and effectively capture global semantic information, enhancing the overall predictive performance of our model.

- **Alternating Attention.** We employ multivariate signals to predict single-channel power series in power time-series forecasting[Das *et al.*, 2023]. Unlike the prevalent transformer-based models that perform feature extraction based on query-key connections among tokens, alternating attention utilizes clustering centers as attention computation cores. This unique approach considers both the Euclidean distance and feature distance between tokens. After nonlinear propagation through activation functions, we implement sequence-level attention filtering, instead of token-level attention, thereby preserving the global characteristics of tokens within each channel to the maximum extent.

SCAT attained strong performance in time-series prediction through these two innovative designs. Existing time-series forecasting methods often adopt sequence decomposition[Xiao *et al.*, 2023] or short-sequence self-learning modeling[Yao *et al.*, 2023] to identify correlations among sub-sequences. However, when applied to power time-series forecasting, numerous indistinguishable anomalies in the sequences lead to a decline in the prediction accuracy[Bonifati *et al.*, 2022]. We propose the SCAT, which incorporates spectral clustering centers to ensure that the model preserves global feature integrity during training. Moreover, we utilize alternating attention as the backbone for feature extraction on series tokens, considering both the Euclidean distance and vector distance perspectives. The implementation of sequence-level attention scores effectively mitigated token oversampling within each channel. Our proposed SCAT model achieved good results through extensive evaluation of multiple prediction datasets.

2 Related Work

This paper's research scope encompasses three primary facets: (i) the investigation of time series clustering; (ii) the development of time series prediction methodologies founded on transformer-based approaches; (iii) the exploration of the utility of alternative architectural models in the domain of power time series prediction.

2.1 Clustering in Time Series Forecasting

As an unsupervised learning algorithm, clustering achieves classification by computing feature distances among entities. K- means [MacQueen, 1967] initiates the process by randomly initializing cluster centers and then partitions samples by optimizing the mean square error within each class. Clustering algorithms that rely on Kullback-Leibler (KL) [Wang and Deng, 2018] divergence excel in spatially partitioning features by minimizing KL loss and bringing samples closer to the cluster center points. In contrast, subspace clustering (SC)[Zhang *et al.*, 2018; Fan *et al.*, 2021] linearly characterizes all samples by partitioning the subspace and optimizing feature representations across different samples. The Gaussian mixture model (GMM) [Peel and MacLahlan, 2000]clustering method assumes that each sample follows an independent Gaussian distribution and addresses the issue of an imbalanced sample distribution by optimizing the likelihood function[Zhang *et al.*, 2021]. In contrast, the clustering method based on mutual information dispenses with any assumptions about the sample distributions. It achieves sample classification by maximizing the mutual information between the input variable, X, and the output variable, Y, effectively mitigating regression problems. The characteristic distribution of power time-series depends on the output curve of power-generation equipment. The clustering based on spectral clustering[Chen *et al.*, 2017; Bianchi *et al.*, 2020] constructs a similarity matrix through sample spacing and uses the topological structure of weather characteristics to classify samples accurately.

2.2 Transformer-Based Time Series Forecasting

The transformer[Vaswani *et al.*, 2017] model offers a more versatile global feature extraction approach compared with the convolutional neural network, and its parallel computing architecture within the attention mechanism has been proved to be faster than the one-way computing method employed by the recurrent neural network. Many research efforts in time-series forecasting have also utilized the transformer architecture. Informer[Zhou *et al.*, 2021] streamlines the calculation time complexity to $O(nlogn)$ through ProbSpare Self-Attention and mitigates error propagation by employing one-step prediction during the forecasting stage. Autoformer further enhances the self-attention mechanism, utilizing Auto-Correlation for subsequence correlation and Series Decomp for time-series decomposition to analyze seasonal information. ETSformer [Woo *et al.*, 2022] adopts an alternative approach, employing multilayered stacked feature extraction layers to unearth the remaining seasonal information embedded in the data's intermediate layers. For longer time-series prediction tasks, FEDformer [Zhou *et al.*, 2022] transforms time-domain information into frequency-domain data and leverages low-rank approximation to deliver advanced performance while significantly reducing computational complexity. Pyraformer [Liu *et al.*, 2022] implements stacked pyramid-shaped attention mechanisms for feature extraction to balance computational efficiency and spatial complexity. Reformer[Kitaev *et al.*, 2020] enhances the computational architecture of attention by removing the query matrix from the traditional attention mechanism and replacing attention weights with functions of keys, thus significantly expanding the sequence length it can accommodate. Patching mechanism-based design, as seen in PatchTST[Nie *et al.*, 2023], has achieved good results in time-series prediction tasks. This mechanism uses subsequences as tokens for feature extraction calculations, ensuring sequence locality. However, in power time-series tasks, the pronounced sequence locality often drifts the entire prediction owing to noise. To address this concern, we introduce the clustering center as the attention computing component and employ a novel token feature-extraction mechanism. This approach effectively balances local developmental trends with global features and optimizes prediction outcomes.

2.3 Other Power Time-Series Forecasting Models

ARIMA[Ariyo *et al.*, 2014], a conventional regression variable algorithm, enjoys widespread usage in wind and solar power prediction. Simultaneously, the evolution of deep learning, particularly long short-term memory (LSTM)[Graves and Graves, 2012; Salinas *et al.*, 2020; Shih *et al.*, 2019], has made significant strides. LSTM's sequential computing mechanism aligns seamlessly with the structural characteristics of time-series flow data, yielding noteworthy research outcomes. Furthermore, equal-recurrent neural networks have garnered considerable attention and research results owing to their suitability for time-series analysis. In specific time intervals, hybrid models like ARMI-ANN[Panigrahi *et al.*, 2018], which fuse ARIMA and artificial neural networks, have demonstrated enhanced predictive capabilities. Additionally, temporal convolutional network (TCN)[Bai *et al.*, 2018] has reduced errors in wind power prediction through its innovative cavity convolution design. Moreover, the combination of generative adversarial networks[Yoon *et al.*, 2019; Luo *et al.*, 2018; Shen *et al.*, 2018] and convolutional networks[Rushe and Mac Namee, 2019] has proved to be effective in improving hourly photovoltaic power predictions. Moreover, some researchers have explored the efficacy of simple linear layer models for time-series prediction, achieving impressive outcomes. DLinear[Zeng *et al.*, 2023], for instance, deconstructs the original data into moving average classification and seasonal trend components. Subsequently, it routes these data through an independent channel of a linear layer to execute series predictions. NLinear, in contrast, enhances prediction results on unevenly distributed datasets through moving averages. SCAT, our time-series prediction algorithm based on the Transformer architecture, leverages alternating attention and centers around the cluster center. This unique approach endows SCAT with a potent combination of deep learning network generalization capability and a linear layer network's intuitive feature extraction prowess. Consequently, it performs better when predicting power-related time-series data.

3 Our Method

This section provides a comprehensive exposition of SCAT, including its operational principles and architectural design. In Section 3.1, we outline the formulation for power time-series prediction. Section 3.2 delves into the intricacies of data preprocessing, while Section 3.3 offers an in-depth exploration of the SCAT architecture. Furthermore, Sections 3.4–3.6 sequentially detail the underlying design principles and calculation formulas for each constituent element within the SCAT framework. The detailed symbol introduction is shown in Table 1.

Notation	Meaning
X	Weather feature at a historical moment
Y	Predicted future time power value
Q	Query: anchor input information
K	Key: the content information of the sequence
V	Value: self-information of the sequence
C	Cluster: cluster center
L	Input feature-length
N	The dataset capacity
$\hat{Q}$	Transition vector from the query to the cluster center
d	Hidden layer attention dimension
k	Number of clusters
e	Vector embedding
y	The real power value at a certain time
$\hat{y}$	The predicted power value at a certain time

Table 1: Main notation table

3.1 Problem Formulation

The power time series comprises two components: the meteorological feature and the power series. These two elements exhibit a one-to-one correspondence in time. We aimed to predict the output power $Y = \{y_{l+1}, y_{l+2}, ..., y_n\}$, denoted as y_t at time t, based on an input vector x consisting of n-dimensional meteorological features of length L is $X = \{\{x_1^1, x_2^1, ..., x_l^1\}, \{x_1^2, x_2^2, ..., x_l^2\}, ..., \{x_1^n, x_2^n, ..., x_l^n\}\}$. We can abstract this problem as a univariate variable forecasting task for a multivariable time series. Before commencing spectral clustering, meticulous analysis and refinement of the dataset are essential.

3.2 Model Architecture

As shown in Figure 2, the computational process of the SCAT involves two key phases: cluster center identification and power prediction. Initially, the model determines the dataset's optimal cluster number and center using the spectral clustering algorithm applied to the input multivariable sequence. The model embeds this cluster center through the cluster center embedding into the subsequent alternating transformer computation architecture as a prompting feature.

In the query projection encoder layer of the alternating transformer, SCAT identifies the nearest cluster center to the current token using a matrix calculation derived from both the cluster center and token features. This utilization of the clustering center enhances the model's feature perception across the entire sequence space. Simultaneously, this approach prevents the attention calculation mechanism from being trapped in local optimal solutions while expanding the model's global perspective. Finally, the final predicted power is output through the linear decoding layer.

3.3 Spectral Clustering

We utilize a Gaussian kernel function in the spectral clustering of SCAT on power time series, as illustrated in the left

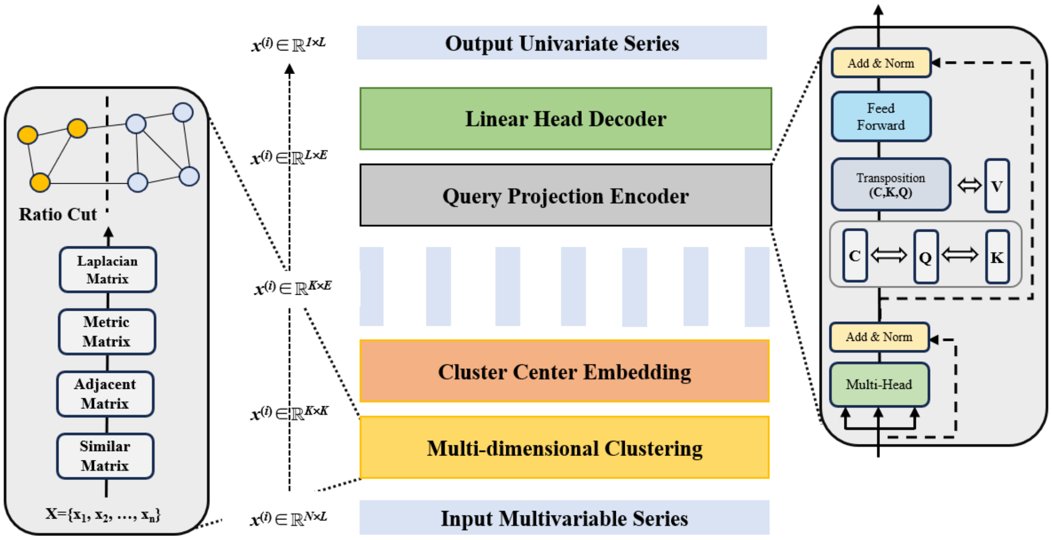

Figure 2: The SCAT architecture consists of three segments: the left segment illustrates the application of spectral clustering for token classification in time series. The middle segment provides an overview of the entire SCAT process. The right segment elaborates on the alternating transformer, which is the fundamental computational mechanism of SCAT.

section of Figure 2. The process begins with calculating the similarity matrix S for the input samples. Subsequently, the adjacency matrix W and the degree matrix D are constructed from S.

$$w_{ij} = s_{ji} = exp(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}) \tag{1}$$

Once the clustering dimension k has been determined, the Laplacian matrix L is computed and normalized.

$$L = D - W \tag{2}$$

Using the normalized Laplacian matrix, we calculate the eigenvector f corresponding to the smallest k eigenvalues. f is then further normalized to create the $[n * k]$-dimensional eigenmatrix F.

$$RatioCut(c1, c2, ...ck) = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 \tag{3}$$

Each row in F is treated as a k-dimensional sample, resulting in n samples. We apply the selected clustering method to these samples, resulting in a clustering dimension of k. Finally, this process yields cluster partition divisions denoted as $C = (c_1, c_2, ..., c_k)$.

To attain optimal spectral clustering outcomes across datasets with varied distributions, we compute the inter-cluster distance to the intra-cluster distance ratio, manipulating multiple gamma values and category numbers, k. The matrix B_k represents the covariance between classes, while W_k represents the covariance matrix within the class. The specific formula is presented below:

$$\begin{aligned} B_k &= \sum_{q=1}^{k} n_q(c_q - c_e)(c_q - C_e)^\top \\ W_k &= \sum_{q=1}^{k}\sum_{x \in C_q}(x - c_q)(x - c_q)^\top \end{aligned} \tag{4}$$

The variable c_q denotes the central point of class q, while c_e represents the central point of the dataset. Additionally, n_q represents the number of data points in class q, and C_q denotes the dataset corresponding to class q.

$$CH = \frac{tr(B_k)(N - K)}{tr(W_k)(K - 1)} \tag{5}$$

The Calinski-Harabasz (CH) metric guided our evaluation, where a higher CH score signifies reduced intra-category covariance, increased inter-category covariance, and enhanced overall classification. We automatically chose the parameters that yielded the most favorable outcomes to obtain the clustering results.

3.4 Alternating Transformers

The spectral clustering algorithm divides tokens into k dimensions based on time points, resulting in the clustering core $[k * k]$ and transforming $[k * k]$ to $[k * e]$ by one-dimensional convolution in the cluster center embedding. Upon acquiring the cluster center, SCAT computes the input sequence matrix to derive the corresponding Q, $\hat{Q}$, and K. These components are then fed into the alternating transformer.

Q undergoes reduction via two convolutional layers to yield an intermediate feature in the computation with $\hat{Q}$. In this case, the feature dimensions of $\hat{Q}$ and C are consistent.

$$Clu\_scores = \sum_{e} \hat{Q}_{hk} C_{kh}^{\top} \tag{6}$$

Figure 3 shows the computational process of the alternating transformer, where $Q = \{q_1, q_2, ..., q_L\}$, $\hat{Q} = \{\hat{q}_1, \hat{q}_2, ..., \hat{q}_L\}$, $C = \{c_1, c_2, ..., c_L\}$. To capture spatial and semantic hierarchical features, we compute the attention values between cluster centers (C) and Q, along with the Euclidean distance between C and $\hat{Q}$. We combine these values and apply the sigmoid activation function for nonlinear transformation. This process yields an eigenmatrix representing tokens' mutual and ownership relationships with cluster centers.

$$Euclidean\_dist(Q, C) = \sqrt{\sum_{i=1}^{n} (Q_i - C_i)^2} \tag{7}$$

At this stage, we consider the affiliation relationship between tokens and cluster centers as determined.

$$Attention\_M = sigmod(-\frac{1}{k}(euclidean\_dist \cdot clu\_scores)) \tag{8}$$

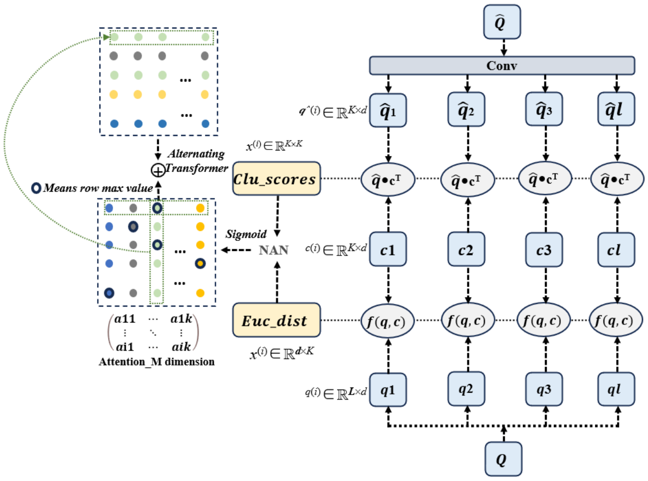
$\hat{Q}$
Conv
$q^{\hat{}}(i) \in \mathbb{R}^{K \times d}$
$\hat{q}_1$
$\hat{q}_2$
$\hat{q}_3$
$\hat{q}l$
$x^{(i)} \in \mathbb{R}^{K \times K}$
Alternating
Transformer
Means row max value
Clu_scores
$\hat{q} \bullet c^{T}$
Sigmoid
NAN
$c(i) \in \mathbb{R}^{K \times d}$
c1
c2
c3
cl
Euc_dist
$f(q,c)$
$x^{(i)} \in \mathbb{R}^{d \times K}$
$q(i) \in \mathbb{R}^{L \times d}$
q1
q2
q3
ql
Q
$\begin{pmatrix} a11 & \cdots & a1k \\ \vdots & \ddots & \vdots \\ ai1 & \cdots & aik \end{pmatrix}$
Attention_M dimension

Figure 3: Vector $\hat{Q}$ undergoes dimensional transformation via the convolution layer and interacts with the cluster center vector C. Then, the Euclidean distance between vectors Q and C is computed. Following the sigmoid transformation between clu_score and euc_dist, the resulting matrix, is denoted as $M = m_{ij}$. Subsequently, this matrix is traversed row-wise, and the maximum value within each row is selected. The column index corresponding to this maximum value is labeled as h. Then, the column eigenvalue $m_i \in M^h$ is transposed, utilizing the column index as the row vector in the output eigenmatrix.

Consequently, we select the I-dimensional feature of each token belonging to the $i - th$ cluster center instead of merging tokens with all cluster centers for vector representation. The Alternating attention matrix is alternately reconstructed using sequence-level features by indexing. The output prediction result is characterized using the softmax function and scaled accordingly. X represents the calculated matrix $[a_{ik}]$ obtained in Figure 3:

$$Alternating = Attention\_M * X_{max\_index}^{\top} \tag{9}$$

This attention computation method transforms token-level features into sequence-level features, preserving the positional characteristics of tokens in the global context and effectively mitigating data anomalies resulting from local fluctuations.

3.5 Loss Function

We employ trade-loss as our loss function, which enhances the loss signal at abrupt points in the time series. $\hat{y} = \hat{y}_i - \hat{y}_{i-1}$ and $y = y_i - y_{i-1}$, Trade loss achieves the maximum extreme value at $Condition = (\hat{y}_i - \hat{y}_{i-1}) * (y_i - y_{i-1}) < 0$.

$$Trade = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} (\hat{y} - y)^2, & Condition < 0 \\ 0, & otherwise. \end{cases} \tag{10}$$

Trade loss as the loss function enables the model to assess the series' trend development state more effectively. In Equation 10, y_i denotes the true value, and $\hat{y}_i$ represents the predicted value.

$$Trade\_loss = \alpha * \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| + (1 - \alpha) * Trade. \tag{11}$$

4 Experiments

Table 2: Detailed description of datasets. Power denotes the maximum power in the wind power time series dataset. Dim denotes the variate number of each dataset. Dataset Size respectively denotes the total number of time points in (Train, Validation, and Test) split.

Datasets	Power	Dim	Dataset Size	Information
Windm1	150 KW	14	(23740, 3391, 6783)	Power
Windm2	49.5KW	14	(23134,3304,6609)	Power
Windm3	100 KW	14	(26604, 3800, 7600)	Power
Windm4	100 KW	14	(18410, 2630, 5260)	Power
Windm5	150 KW	14	(26446, 3778, 7556)	Power
ETTm1	-	7	(34465, 11521, 11521)	Electricity
ETTm2	-	7	(34465, 11521, 11521)	Electricity
Weather	-	21	(36792, 5271, 10540)	Weather
Traffic	-	862	(12185, 1757, 3509)	Transportation
Electricity	-	321	(18413, 2630, 5261)	Electricity

4.1 Datasets

We extensively examined 10 real-world datasets in our experiments. (1) ETT encompasses power transformer data with seven variables over 2 years, from July 2016 to July 2018, collected at 15-minute intervals. (2) Weather includes weather data spanning 21 variables in 2020, recorded at 10-minute intervals. (3) Traffic consisted of traffic data across 862 variables gathered between January 2015 and December 2016. (4) Electricity includes 321 variables of electricity data

Methods	SCAT(Our)		DLinear		PatchTST		FEDformer		Pyraformer		LightTS		TimesNet		MICN		Autoformer		ETSformer		HI		STID	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Windm1	**0.516**	**0.481**	0.687	0.650	0.623	0.550	0.567	0.569	0.650	0.600	0.652	0.609	0.620	0.527	0.624	0.599	0.822	0.687	0.671	0.607	1.027	0.711	1.076	0.855
Windm2	0.614	**0.558**	**0.596**	0.646	0.644	0.599	0.624	0.616	0.610	0.620	0.596	0.629	0.628	0.578	0.562	0.597	0.825	0.702	0.617	0.592	1.036	0.741	0.830	0.802
Windm3	0.571	**0.503**	0.601	0.647	0.662	0.552	0.652	0.621	0.567	0.594	0.583	0.610	0.685	0.557	**0.544**	0.579	1.005	0.776	0.673	0.637	0.149	0.736	0.881	0.833
Windm4	**0.589**	**0.512**	0.605	0.640	0.639	0.553	0.661	0.577	0.589	0.600	0.611	0.626	0.638	0.534	0.590	0.603	0.873	0.690	0.647	0.602	1.086	0.708	0.811	0.775
Windm5	0.754	**0.587**	0.699	0.677	0.793	0.609	0.842	0.662	0.677	0.639	0.698	0.656	0.870	0.627	**0.641**	0.616	1.091	0.772	0.885	0.762	1.458	0.836	0.979	0.864
ETTm1	0.367	0.486	0.040	0.143	**0.034**	**0.134**	0.043	0.158	0.133	0.291	0.101	0.243	0.033	0.134	0.043	0.156	0.053	0.178	0.204	0.368	0.057	0.184	2.293	1.350
ETTm2	0.495	0.528	0.101	0.240	**0.078**	**0.196**	0.083	0.217	0.331	0.408	0.113	0.245	0.088	0.213	0.082	0.210	0.135	0.286	0.196	0.345	0.265	0.392	1.782	1.239
Electricity	0.283	0.404	0.382	0.465	**0.271**	**0.381**	0.362	0.472	0.335	0.443	0.353	0.453	0.260	0.360	0.241	0.367	0.346	0.453	0.554	0.579	1.006	0.744	0.948	0.800
Traffic	0.341	0.406	0.449	0.479	0.188	0.274	0.233	0.331	0.381	0.420	0.605	0.585	**0.175**	**0.271**	0.207	0.295	0.292	0.391	1.313	0.907	2.345	1.036	0.627	0.587
Weather	**0.001**	**0.028**	0.004	0.049	0.004	0.049	0.020	0.088	0.002	0.041	0.003	0.041	0.004	0.047	0.008	0.055	0.023	0.094	0.082	0.187	0.005	0.056	0.004	0.043

Table 3: The average of multivariate-input-univariate-output forecasting results with prediction lengths $S \in \{32, 64, 96, 192, 336\}$ for all methods.

gathered between July 2016 and July 2019, recorded at hourly intervals. (5) Windm comprises five wind power-generation datasets from various regions, incorporating 14 pertinent features, collected at 15-minute intervals. The details of datasets are presented in Table 2.

4.2 Forward process

Before conducting spectral clustering, we had to thoroughly analyze and refine the dataset. This refinement process encompassed two main aspects: handling missing points and addressing outliers. We used an interpolation method employing the average of neighboring critical values to address discrete missing data points. Conversely, we considered large sections with substantial minus points. Notably, values exceeding these limits were adjusted to correspond with the accepted maximum wind speed to account for differing operational parameters regarding the maximum wind speed across stations. Additionally, data normalization was conducted, particularly for attributes with disparate dimensions, such as wind direction and wind speed. Considering the directional nature of wind (ranging from 0°to 360°, we achieved normalization using the transformation formula (x - 0.00001) / 360.

4.3 Experimental Settings

The experimental settings are presented in baselines, evaluation criteria, and detailed experimental parameter settings.

Baselines: Our model selection process focused on contemporary models that have demonstrated good performance in recent time series prediction tasks. The 11 baseline models encompassed Autoformer[Wu *et al.*, 2021], PatchTST [Nie *et al.*, 2023], DLinear[Zeng *et al.*, 2023], ETSformer[Woo *et al.*, 2022], FEDformer[Zhou *et al.*, 2022], Pyraformer [Liu *et al.*, 2022], LightTS[Campos *et al.*, 2023], MICN[Wang *et al.*, 2023]and TimesNet[Wu *et al.*, 2023]. Additionally, we included two models, HI[Cui *et al.*, 2021] and STID[Shao *et al.*, 2022], tailored explicitly for power time series prediction.

Evaluation Metrics: We employed the mean square error (MSE) and mean absolute error (MAE) as fundamental metrics for comparison.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (12)$$

Detailed Settings: The SCAT architecture includes two encoder layers and one decoder layer. We determined the optimal classification number for diverse datasets by combining gamma and k parameters through spectral clustering. We set the cluster center dimension to k, fixed the model dimension at 512, and used eight attention components. Additionally, we applied a dropout rate of 0.1. To mitigate randomness, we conducted three repeated experiments for each experimental group, determining the model's final performance based on the best and average results. The training batch size was 32. We utilized Adam with an initial learning rate in $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}\}$. The experimental framework used Pytorch and the algorithm ran on RTX 3080 GPU.

4.4 Results and Analysis

Table 3 presents the combined performance of the multivariate forecast univariate tasks of SCAT and the baseline models across 10 datasets. encompassing five wind power datasets. SCAT achieved good results on the Wind and weather datasets, yielding an average comprehensive prediction improvement of 10.2% and 49.1% respectively compared with DLinear. Notably, on the Wind dataset, SCAT's performance notably surpassed that of specifically designed models like HI and STID for power time-series, showcasing an overall performance enhancement of 40%. The correlation between wind power and meteorological factors contributed to SCAT's performance increase in the weather dataset compared with PatchTST's suboptimal performance. However, SCAT's sensitivity to highly fluctuating data distributions led to numerous ineffective feature calculations within its alternating transformer mechanism for ETT and electricity datasets, resulting in degraded prediction performance due to their smoother data distributions.

In long-term power-series prediction, the model utilizes future meteorological features obtained at present, leading to continuous errors accumulating during the series decomposition and inference processes. These errors escalated with longer forecast windows, causing rapid increases in forecast losses. SCAT excelled over long durations owing to the inherent spatial information stored in the cluster core, providing accurate spatial features consistently. Among the baselines, PatchTST performed most similarly to SCAT. This proximity in performance can be attributed to these datasets' shared concept of sequence-level learning, allowing for more comprehensive acquisition of semantic information and closer prediction of actual power values.

DLinear, employing a superficial linear layer for feature extraction, abandons complex methods, achieving enhanced prediction performance. However, in the visual prediction results, wind power remained zero for approximately 6 hours daily owing to insufficient wind speed, leading to continuous zeros in the power sequence. These actual but continuous zero values cannot be considered outliers. Nevertheless, in DLinear, these zeros significantly impacted the fully connected layer's calculations, causing substantial deviations in the prediction results.

Training Efficiency

We assessed the training efficiency of all models under identical conditions, and the results are presented in Figure 4. Notably, the linear predictive models DLinear and LightTS exhibited the shortest training times for a single epoch. In transformer-based models, ETSformer demonstrated the highest efficiency, followed by SCAT, which outperformed Pyraformer, Autoformer, and FEDformer in training efficiency. Conversely, TimesNet, relying on graph information mining, exhibited the slowest training efficiency among all models. It required almost four times the computational time of SCAT. In summary, SCAT showcased slightly superior computational efficiency compared with other transformer-based models.

5 Ablation Study

We comprehensively assessed each component's role in SCAT through ablation experiments, as shown in Fig. 5. SCAT's primary objective is to utilize cluster centers to influence global features in sequences of varying lengths. To evaluate this, we achieved it by randomly initializing cluster centers or setting them as zero vectors to eliminate global features' impact on SCAT deliberately. However, the random initialization strategy had a more significant negative impact. Simultaneously, we assessed the impact of the loss function on SCAT's predictive performance. The MSE indicator employing trade-loss exhibited a decrease across all datasets. Because trade-loss bolstered the sequence difference loss, it led to an attenuation in optimizing MAE. However, regarding comprehensive performance, SCAT using trade-loss showed lower predictive losses.

6 Conclusion and Future Work

This paper introduced the SCAT model for time-series prediction, incorporating two significant innovations. First, it incorporates clustering centers as global covariables to adjust the model's prediction preferences in long-distance series predictions. Addressing the challenge of falling into local optimal solutions significantly enhances prediction accuracy during the prediction process. Second, the token-level feature selection is elevated to the sentence level by replacing the original full attention with an alternating attention mechanism. Additionally, two characteristic calculation methods, Euclidean distance and token similarity in numerical space, mitigate unreasonable calculation results in time-series prediction. However, during spectral clustering, the clustering logic of the algorithm is not aligned with the actual power numerical distribution, leading to clustering centers being biased toward theoretical centers rather than actual centers. In future research, we aim to explore the practical application of clustering algorithms in power time-series. We will focus on developing a clustering algorithm closely aligned with actual power distribution to enhance SCAT's prediction performance further.

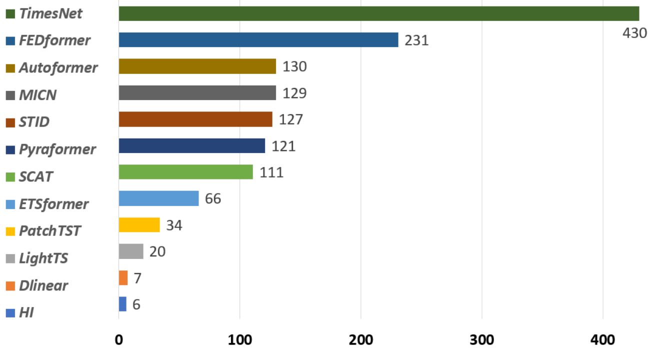

Figure 4: Comparing the training time of a single epoch of SCAT and other baselines. Abscissa represents model training time.

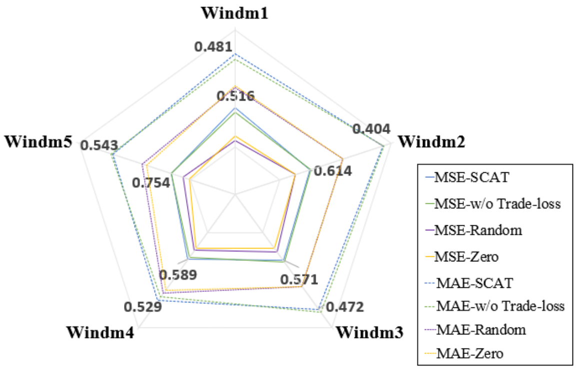

Figure 5: The ablation test examines the SCAT components. 'w/o Trade-loss' signifies SCAT employing MSE as a loss function. 'Random' refers to utilizing a randomly initialized vector as the cluster center. 'Zero' involves an all-0 vector as the cluster center. For more precise visualization, dashed lines represent MAE, solid lines represent MSE, and the graphical scale of MSE was adjusted.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No.2021ZD0112400), the 111 Project(No.D23006), the National Natural Science Foundation of China (Nos.62076045, 62202080), the Basic Project for Universities from the Educational Department of Liaoning Province(No.LJKFZ20220290), the Science and Technology Project of Liaoning Province (No.2023JH1/10400083), Dalian Major Projects of Basic Research (No.2023JJ11CG002).

References

[Ariyo *et al.*, 2014] Adebiyi A Ariyo, Adewumi O Adewumi, and Charles K Ayo. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*, pages 106–112. IEEE, 2014.

[Bai *et al.*, 2018] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*, 2018.

[Bianchi *et al.*, 2020] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International conference on machine learning*, pages 874–883. PMLR, 2020.

[Bonifati *et al.*, 2022] Angela Bonifati, Francesco Del Buono, Francesco Guerra, and Donato Tiano. Time2feat: learning interpretable representations for multivariate time series clustering. *Proceedings of the VLDB Endowment*, 16(2):193–201, 2022.

[Box *et al.*, 2015] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[Campos *et al.*, 2023] David Campos, Miao Zhang, Bin Yang, Tung Kieu, Chenjuan Guo, and Christian S. Jensen. LightTS: Lightweight time series classification with adaptive ensemble distillation. *Proceedings of the ACM on Management of Data*, 1(2):1–27, jun 2023.

[Chan *et al.*, 2015] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015.

[Chen *et al.*, 2017] Dongdong Chen, Jiancheng Lv, and Yi Zhang. Unsupervised multi-manifold clustering by learning deep representation. In *Workshops at the thirty-first AAAI conference on artificial intelligence*, 2017.

[Cui *et al.*, 2021] Yue Cui, Jiandong Xie, and Kai Zheng. Historical inertia: A neglected but powerful baseline for long sequence time-series forecasting. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 2965–2969, 2021.

[Das *et al.*, 2023] Abhimanyu Das, Weihao Kong, Andrew Leach, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Doubleday *et al.*, 2020] Kate Doubleday, Stephen Jascourt, William Kleiber, and Bri-Mathias Hodge. Probabilistic solar power forecasting using bayesian model averaging. *IEEE Transactions on Sustainable Energy*, 12(1):325–337, 2020.

[Elsaraiti and Merabet, 2022] Meftah Elsaraiti and Adel Merabet. Solar power forecasting using deep learning techniques. *IEEE Access*, 10:31692–31698, 2022.

[Fan *et al.*, 2021] Lili Fan, Guifu Lu, Yong Wang, and Tao Liu. Block diagonal sparse subspace clustering. In *2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6, 2021.

[Graves and Graves, 2012] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[Khalifa *et al.*, 2022] Ahmed A Khalifa, Abdul-Jalil Ibrahim, Abdulkarem I Amhamed, and Muftah H El-Naas. Accelerating the transition to a circular economy for net-zero emissions by 2050: a systematic review. *Sustainability*, 14(18):11656, 2022.

[Kitaev *et al.*, 2020] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

[Le Guen and Thome, 2020] Vincent Le Guen and Nicolas Thome. Probabilistic time series forecasting with shape and temporal diversity. *Advances in Neural Information Processing Systems*, 33:4427–4440, 2020.

[Lim and Zohren, 2021] Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

[Liu *et al.*, 2022] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*, 2022.

[Luo *et al.*, 2018] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. Multivariate time series imputation with generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

[MacQueen, 1967] James MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(14):281–297, 1967.

[Moradzadeh *et al.*, 2021] Arash Moradzadeh, Hamed Moayyed, Sahar Zakeri, Behnam Mohammadi-Ivatloo, and A Pedro Aguiar. Deep learning-assisted short-term load forecasting for sustainable management of energy in microgrid. *Inventions*, 6(1):15, 2021.

[Nie *et al.*, 2023] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

[Panigrahi *et al.*, 2018] Sibarama Panigrahi, Himansu Sekhar Behera, and Ajith Abraham. A fuzzy filter based hybrid arima-ann model for time series forecasting. In *Proceedings of the Eighth International*

Conference on Soft Computing and Pattern Recognition (SoCPaR 2016), pages 592–601. Springer, 2018.

[Peel and MacLahlan, 2000] DAVID Peel and G MacLahlan. Finite mixture models. *John & Sons*, 2000.

[Rushe and Mac Namee, 2019] Ellen Rushe and Brian Mac Namee. Anomaly detection in raw audio using deep autoregressive networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3597–3601. IEEE, 2019.

[Salinas *et al.*, 2020] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

[Shao *et al.*, 2022] Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 4454–4458, New York, NY, USA, 2022. Association for Computing Machinery.

[Shen *et al.*, 2018] Zhipeng Shen, Yuanming Zhang, Jiawei Lu, Jun Xu, and Gang Xiao. Seriesnet: a generative time series forecasting model. In *2018 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2018.

[Shih *et al.*, 2019] Shun-Yao Shih, Fan-Keng Sun, and Hung-yi Lee. Temporal pattern attention for multivariate time series forecasting. *Machine Learning*, 108:1421–1441, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[Voyant *et al.*, 2017] Cyril Voyant, Gilles Notton, Soteris Kalogirou, Marie-Laure Nivet, Christophe Paoli, Fabrice Motte, and Alexis Fouilloy. Machine learning methods for solar radiation forecasting: A review. *Renewable energy*, 105:569–582, 2017.

[Wang and Deng, 2018] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[Wang *et al.*, 2011] Xiaochen Wang, Peng Guo, and Xiaobin Huang. A review of wind power forecasting models. *Energy procedia*, 12:770–778, 2011.

[Wang *et al.*, 2023] Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. Micn: Multi-scale local and global context modeling for long-term series forecasting. In *International Conference on Learning Representations*, 2023.

[Woo *et al.*, 2022] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven C. H. Hoi. Etsformer: Exponential smoothing transformers for time-series forecasting. In *International Conference on Machine Learning*, 2022.

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

[Wu *et al.*, 2023] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

[Xiao *et al.*, 2023] Fei Xiao, Yuncheng Wu, Meihui Zhang, Gang Chen, and Beng Chin Ooi. Mint: Detecting fraudulent behaviors from time-series relational data. *Proceedings of the VLDB Endowment*, 16(12):3610–3623, 2023.

[Yao *et al.*, 2023] Yuanyuan Yao, Dimeng Li, Hailiang Jie, Hailiang Jie, Tianyi Li, Jie Chen, Jiaqi Wang, Feifei Li, and Yunjun Gao. Simplets: An efficient and universal model selection framework for time series forecasting. *Proceedings of the VLDB Endowment*, 16(12):3741–3753, 2023.

[Yoon *et al.*, 2019] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

[Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, Jun. 2023.

[Zhang *et al.*, 2018] Zheng Zhang, Yong Xu, Ling Shao, and Jian Yang. Discriminative block-diagonal representation learning for image recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 29(7):3111–3125, 2018.

[Zhang *et al.*, 2021] Dejiao Zhang, Feng Nan, Xiaokai Wei, Shangwen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew Arnold, and Bing Xiang. Supporting clustering with contrastive learning, 2021.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11106–11115, 2021.

[Zhou *et al.*, 2022] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286. PMLR, 2022.