HALO: HVAC Load Forecasting with Industrial IoT and Local-Global-Scale Transformer

Cheng Pan, *Student Member, IEEE*, Cong Zhang, Edith C. H. Ngai, *Senior Member, IEEE*, Jiangchuan Liu, *Fellow, IEEE*, and Bo Li, *Fellow, IEEE*

***Abstract*—The evolution of Internet-of-Things (IoT) is fostering the use of intelligent controls for energy conservation. Yet, the efficacy of these strategies is largely tied to diverse load forecasting algorithms. Given the significant contribution of heating, ventilation, and air-conditioning (HVAC) systems to global energy consumption, accurate forecasting of HVAC power usage is crucial for improving overall energy efficiency. However, real-world HVAC load forecasting, bolstered by various IoT devices, is complicated by multiple factors: data variability, power load fluctuations, electronic phenomena (e.g., zero drifts), and the increased time complexity and larger model sizes required to manage accumulating historical data. To address these challenges, we first present an in-depth measurement study on the characteristics of HVAC load at a minute scale based on HVAC data collected in six locations. We propose HALO, a transformer-based framework specifically designed for forecasting HVAC load. HALO incorporates an adaptive data pre-processing stage and a local-global-scale transformer-based load forecasting stage, enabling precise forecasting of HVAC load and optimization of energy utilization. Evaluation based on real-world data traces from a prototype application demonstrates that the proposed framework significantly outperforms existing models.**

***Index Terms*—Internet of Things (IoT), Smart Energy, Energy conservation, Load forecasting, Transformer**

I. Introduction

In order to adhere to the 1.5°C target set in the Paris Agreement, emissions must be decreased by 45% by 2030 and achieve carbon neutrality by 2050 [1]. Among various sectors, building energy consumption holds a substantial share in overall energy usage, accounting for 30% of global energy consumption [2]. Within building energy consumption, the operation of Heating, Ventilation, and Air-Conditioning (HVAC) systems plays a crucial role, representing approximately 40% - 60% of total energy consumption in buildings [3]. As the Internet-of-Things (IoT) evolves, it enables intelligent control of HVAC systems, leading to substantial advancements in energy conservation [4]. However, the effectiveness of these intelligent HVAC control strategies hinges largely on the application of accurate and timely load forecasting algorithms [5].

Accurate and timely HVAC load prediction enables building control systems to operate more efficiently by matching energy usage with actual air-conditional demand, thereby reducing power consumption and further lowering carbon emissions [6, 7, 8, 9]. Recently, This has sparked significant interest among researchers in both industry and academia, focusing on HVAC load forecasting to optimize energy efficiency by adjusting HVAC operations. [10]. HVAC load forecasting methods can be broadly classified into two main categories: *physical-based* models and *data-driven* models. Physical-based models, also known as white-box models, rely on fundamental physical principles to describe the heat transfer characteristics of buildings [11]. However, these models require extensive and detailed building information, and their prediction accuracy may vary if the underlying assumptions of the physical principles are not consistently met [12]. On the other hand, data-driven models, referred to as black-box models, leverage shallow machine learning or deep learning techniques, offering distinct advantages for building energy load forecasting [10].

To understand their effectiveness and investigate the challenges in real-world HVAC data, we have developed EdgeSpot, an edge computing-based IoT device specifically designed for smart energy, as shown in Figure 1. EdgeSpot serves as an IoT hub for communication and computing, enabling seamless interaction, minute-level data acquisition, and processing with diverse IoT devices, including electricity meters, HVAC communication panels, and meteorological instruments. We have collected measurements and made observations on diverse HVAC systems across different buildings, user behaviors, and locations. The preliminary data presented several significant challenges. First, the data exhibits variability due to geographic differences, diverse HVAC system brands and types, and varying numbers of internal units. This complexity poses difficulties in accurate modeling and analysis. Second, power load fluctuates in response to different features across various time frames. For example, temperature strongly influences weekly load patterns but has less impact on shorter time windows (e.g., 24 hours, 1 hour, or 15 minutes). On the other hand, the count of active HVAC internal units, reflecting user behaviors, significantly affects load even within a 15-minute window. Third, we noticed two electrical phenomena in real-world data: zero drifts and voltage spikes. These phenomena can be caused by factors such as temperature and humidity changes, power surges, or switch-tripping on IoT devices. As a result, existing

Cheng Pan is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: cpanpan@connect.hku.hk)

Cong Zhang is with Jiangxing Intelligence Inc. and the Department of Computer Science, The University of Hong Kong, Hong Kong (e-mail: zhangcong@jiangxingai.com)

Edith C. H. Ngai is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: chngai@eee.hku.hk)

Jiangchuan Liu is with the School of Computing Science, Simon Fraser University, British Columbia, Canada. (e-mail: jcliu@sfu.ca)

Bo Li is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong (e-mail: bli@ust.hk)

Corresponding authors: Cong Zhang; Edith C. H. Ngai.

Copyright (c) 20xx IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

models often struggle to accurately predict during transient periods and load fluctuations, leading to practical limitations in real-world applications.

Our observations underline the need for advanced techniques for managing data variability and capturing temporal dynamics. Given the rapid growth of model size and complexity, we advocate the adoption of *transformer-based* models, which have shown considerable promise in building load forecasting tasks [13, 14]. These models, equipped with a *self-attention* mechanism [15], offer several advantages such as reduced complexity, fewer parameters, lower computational requirements, and the ability to capture intricate temporal dependencies. These advantages make them well-suited to address the challenges of larger model sizes and increased time complexity encountered in existing deep learning approaches. However, the current transformer-based models [13, 14] predominantly rely on the vanilla transformer [15], which exhibits limitations, including fixed-length attention and limited sequential dependency. While the vanilla transformer excels at capturing short-term dependencies, but encounters difficulties in handling long-term dependencies and complex patterns. For example, the attention mechanism tends to prioritize nearby time steps, potentially overlooking broader temporal relationships that significantly influence peak values. Consequently, these models may inadequately address the complexities of real-world HVAC load forecasting, such as data variability, power load fluctuations, and electronic phenomena.

Therefore, to address the aforementioned challenges, we propose HALO, a local-global-scale transformer-based framework for HVAC load forecasting. To assess the generalizability of our framework, we conduct experiments evaluating its performance on distinct buildings with different geo-locations and user behaviors. We observe the superiority of the proposed framework in the accuracy and reliability of load forecasting for HVAC systems. This improved forecasting capability contributes to optimizing energy efficiency in HVAC systems, making progress toward achieving net-zero emissions [6, 7, 8]. Our contributions to this work can be summarized as follows:

- We developed EdgeSpot, an edge computing-based IoT equipment, to establish communication with a diverse range of IoT devices, including electricity meters, HVAC communication panels, and meteorological instruments.
- We collects comprehensive data from six buildings with diverse locations and user behaviors by deploying EdgeSpots and environment sensors. The dataset includes information on indoor and outdoor environmental factors, electricity-related data, and specific operating details of HVAC systems. This extensive data collection provides valuable insights into the factors influencing HVAC load, such as data variability, power load fluctuations, and electronic phenomena.
- We propose HALO, a comprehensive transformer-based framework for HVAC load forecasting. HALO consists of three main stages: an adaptive data pre-processing stage, a transformer-based load forecasting stage with multiple encoders, and a scale fusion stage. The adaptive data pre-processing stage addresses challenges related to data variability, zero drift, and voltage spikes, which are brought by IoT devices during the data collection. The transformer-based HVAC load forecasting stage effectively captures long-range global dependencies and local information across different temporal windows. The scale fusion stage captures the information of the original series while emphasizing the forecasting performance for the peak load values.
- By integrating information from global and local encoders, HALO significantly enhances the modeling capability and performance of load forecasting models. Empirical evaluations across different locations demonstrate at least an improvement of 3.13 times in performance on 24-hour load forecasting compared to existing methods.

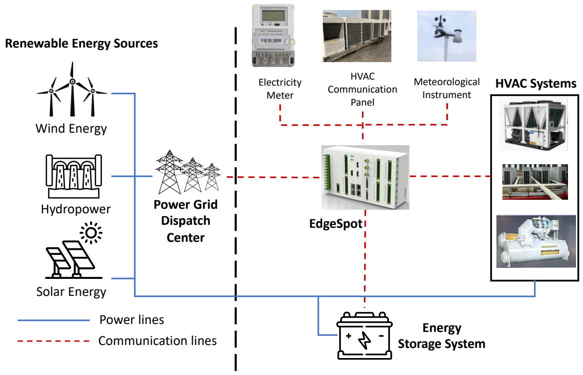


Fig. 1: Deployment diagram of EdgeSpot

II. HVAC Measurements with EdgeSpot

A. Measurement Implementation

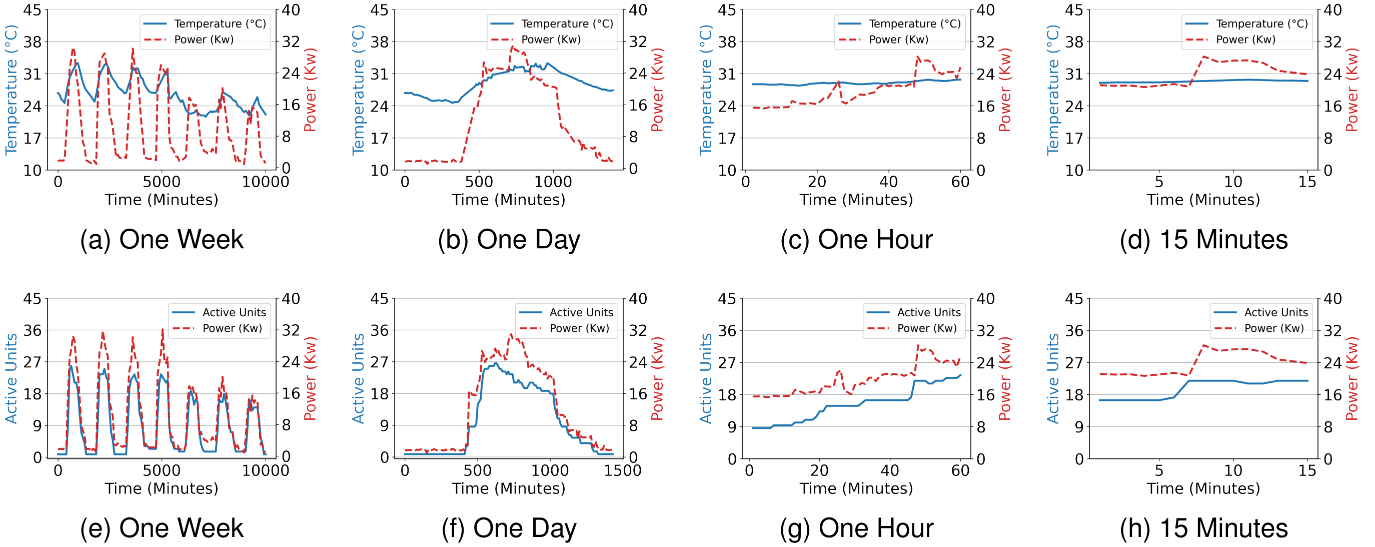
HVAC systems encompass various components such as heating equipment, ventilation equipment, and cooling or air-conditioning equipment, all of which are essential for maintaining comfortable indoor environments in diverse settings like residential, commercial, and industrial buildings. These systems play a crucial role in regulating temperature, humidity, and air quality inside buildings and locations [16].

EdgeSpot. The advent of IoT-based smart devices has led to the development of numerous cost-effective HVAC load forecasting models [5]. These advancements are further enhanced by edge computing-based IoT systems, which enable real-time data processing and the execution of lightweight AI algorithms [17, 18, 19]. In order to better understand the dynamics of HVAC systems and identify real-world data challenges inside, we deployed an edge computing-based IoT equipment, called EdgeSpot, specifically designed for smart energy (see Figure 1). A diverse range of IoT devices, such as electricity meters, HVAC communication panels, and meteorological instruments, can be interfaced with EdgeSpot. As an edge computing-based IoT device, it serves as a crucial intermediary for transferring information between the smart power grid dispatch center [20], energy storage systems [21], and HVAC systems, thereby facilitating better coordination and management of renewable energy resources.

In our study, we have accomplished the integration of EdgeSpot into smart energy systems, thereby facilitating seamless communication with a wide range of IoT devices via

TABLE I: Buildings demonstrate variability due to their geographical locations, environmental factors, and user behaviors.

Building	Type	# of Features	Date Range	# of Internal Units	User Behavior	City Climate
A	4S Dealership (Brand A)	45	7/12/2023-9/11/2023	50	Opening hours	City A: Humid subtropical climate, four distinct seasons, and moderate amount of rainfall
B	4S Dealership (Brand B)	29	8/15/2023-9/13/2023	13	Opening hours	
C	4S Dealership (Brand C)	29	8/18/2023-9/14/2023	33	Opening hours	
D	Airbnb	29	8/18/2023-9/13/2023	35	Occupancy and guest preferences	
E	Office Building	29	5/09/2023-6/1/2023	131	Occupancy and working hours	City B: Subtropical monsoon climate, hot, long and humid summers, and substantial amount of rainfall
F	Office Building	29	5/09/2023-6/2/2023	240	Occupancy and working hours	



(a) One Week (b) One Day (c) One Hour (d) 15 Minutes

(e) One Week (f) One Day (g) One Hour (h) 15 Minutes

Fig. 2: The power load of a VRV system varies in response to changes in temperature and the number of active internal units.

multiple interfaces. EdgeSpot incorporates a CPU of ARM 4 core Cortex-A55 operating at 1.8GHz, complemented by an independent Neural Processing Unit (NPU). This configuration not only satisfies the stringent security prerequisites of smart energy but also delivers optimized computational power. EdgeSpot boasts up to 3GB of available storage and up to 1.5GB of memory. EdgeSpot exhibits the capability to gather millisecond-level data and exercise control at the same granularity. Yet, to meet the storage and measurement requirements, we collect data at a 15-minute granularity. These attributes contribute significantly to regional control by facilitating the deployment of locally tailored control strategies.

B. Data Variability

To better understand the dynamics of HVAC systems, we collected data from six buildings as illustrated in Table I[1] Data from diverse HVAC systems across multiple buildings demonstrate variability due to their geographical locations, environmental factors, and user behaviors. Specifically, Buildings A to D and E to F are located in two separate cities within different provinces of China, with an approximate straight-line distance of 440 kilometers between them. These two provinces exhibit divergent climate patterns. Buildings A to D, located in a region with a subtropical monsoon climate, experience hotter summers and colder winters, while buildings E and F, located in a tropical monsoon climate, enjoy milder, longer summers and shorter winters [22]. Moreover, each building is equipped with a different count of internal units. Each of these units is governed by an HVAC communication panel, which captures specific features relevant to its corresponding area within the building. The presence of such data inconsistencies across different buildings accentuates the need to address the increased count of features and serves as a motivation for applying data pre-processing techniques.

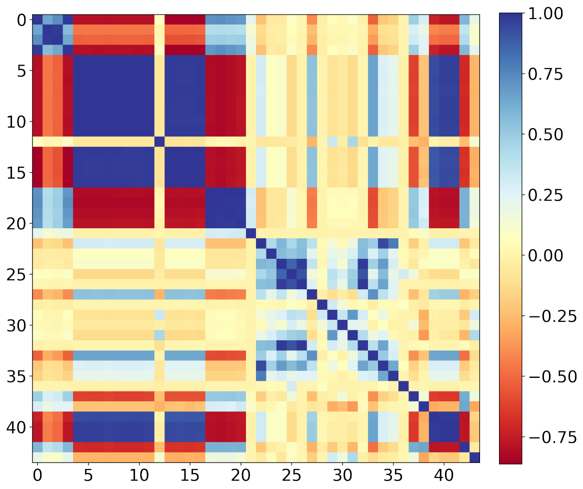


Fig. 3: Feature Correlation for Location A.

[1] According to the data collection agreement with the data providers, we omit certain details about each building.

TABLE II: Collected Features

Electricity Meters			
0	The Power Factor Angle (ϕ)		
1	The Power Factor Angle of Phase a (ϕ_a)		
2	The Power Factor Angle of Phase b (ϕ_b)		
3	The Power Factor Angle of Phase c (ϕ_c)		
4	Total Current (I)		
5	Phase a Current (I_a)		
6	Phase b Current (I_b)		
7	Phase c Current (I_c)		
8	Total Power Load (P)		
9	Phase a Active Power (P_a)		
10	Phase b Active Power (P_b)		
11	Phase c Active Power (P_c)		
12	Total Forward Active Energy (Pev)		
13	Total Reactive Power (Q)		
14	Phase a Reactive Power (Q_a)		
15	Phase b Reactive Power (Q_b)		
16	Phase c Reactive Power (Q_c)		
17	Total Voltage (U)		
18	Phase a Voltage (U_a)		
19	Phase b Voltage (U_b)		
20	Phase c Voltage (U_c)		
21	Operating Frequency (f)		
Meteorological Instruments		**Communication Panels**	
22	Real Time Wind Speed	37	Humidity
23	Real Time Wind Direction	38	Temperature
24	Atmospheric Temperature	39	VRV Error Code
25	Atmospheric Humidity	40	VRV Fan Speed
26	Atmospheric Pressure	41	VRV Mode
27	Illuminance	42	VRV On Off
28	Minute Rainfall	43	Room Temperature
29	Hourly Rainfull	44	VRV Temperature
30	Daily Rainfall		
31	Accumulated Rainfall		
32	Dew Point Temperature		
33	Radiation		
34	Wind Force		
35	Wave Height		
36	Wind Direction Angle		

The features collected from our edge computing-based IoT device, EdgeSpot, are presented in Table II, encompassing a wide range of information. These include data on electrical energy consumption, air conditioning system operation, and user behaviors, as well as real-time meteorological conditions. Specifically, we collected a total of 45 features from building A, which was the only building where we successfully implemented meteorological instruments. For buildings B to F, we collected a total of 29 features. This includes all the features obtained from the electricity meter and HVAC communication panel, with the exception of one feature related to operating frequency.

C. Feature Correlation

To find the feature correlation among the features in our dataset, we conducted a preliminary correlation analysis [23], as shown in Figure 3, we observed significant correlations among the features. Specifically, a strong positive correlation emerged between reactive power Q (Feature 13-15), current I (Feature 4-7), and power load P (Feature 8-11). Yet, the total forward active energy (Feature 12) and reactive power Q, current I, and power load P, although related, are not directly correlated due to their different natures and units of measurement. Additionally, we found a positive correlation between power load P and various features (Feature 39-42) obtained from HVAC communication panels, such as the number of active internal units indicated by the 42nd feature. This suggests that a higher number of active internal units tends to coincide with higher levels of power load P. On the other hand, weather factors (Feature 22-38) demonstrate correlations with power load P, and they show a relative impact on it. These findings shed light on the interrelationships between different features and contribute to a better understanding of the factors influencing power load P. However, these findings do not provide comprehensive insights into how these collected features specifically impact the power load across different time windows. To gain a deeper understanding of the relationship between the collected features and the power load, we next conducted an analysis of the impact of features on the power load within different time windows.

D. Power Load Fluctuations

The fluctuations observed in power load can be attributed to a variety of factors, including weather conditions and user behaviors. These factors influence power load within different time windows.

We collected one week of data from August 24th to 31st. Daily data was gathered specifically on August 24th. For a more granular analysis, we obtained one-hour data from 8-9 a.m. and 15-minute data from 8:40-8:55 a.m. on the same day. We found that the power load exhibited variations in response to changes in specific features across various time windows. Specifically, as shown in Figure 2, we found that temperature was a significant weather factor that impacted the power load. The temperature and power load changes exhibit both weekly and daily patterns.

In the meantime, we found that temperature changes had much less influence on the power load during shorter time windows. In contrast, user behaviors may have a more significant impact on the power load in short time intervals than weather. For example, as shown in Figure 2c and Figure 2d, the power load significantly increased when the dealership opened at approximately 8:30 a.m. In addition, the number of active internal units, a significant feature obtained from HVAC communication panels that reflects user behaviors, clearly impacted the power load within a 15-minute time window. As the dealership opened, users turned on the HVAC systems, leading to an increase in the number of active internal units and a significant increase in the power load. These findings highlight that weather impacts are noticeable over extended periods like a day or week, while user behaviors affect power load within shorter windows, such as an hour or 15 minutes. Weekly and daily patterns provide a comprehensive view and understanding of power load fluctuations. This motivates us to develop a novel model that incorporates suitable time windows, effectively capturing and adapting to information from various time intervals.

E. Zero Shift and Voltage Spike

In our collected data, we also observed two common phenomena that can occur in HVAC systems: *zero drift* and

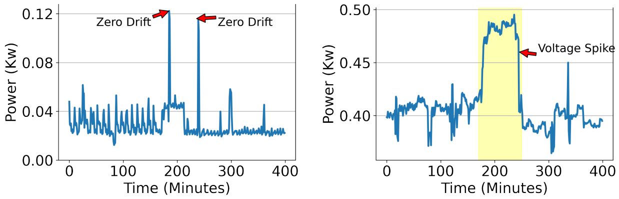


Fig. 4: Left: Zero drifts. Right: Voltage Spikes.

voltage spikes, as illustrated in Figure 4. Zero drift refers to the gradual deviation of a sensor's output from its true value over time, even without any input. This phenomenon is influenced by factors such as temperature and humidity changes, or switch tripping in IoT devices, resulting in a shift in measured values [24]. On the other hand, voltage spikes are sudden and transient increases in electrical voltage caused by events like lightning strikes, power surges, or high-powered device switching on IoT devices [25]. Both zero drift and voltage spikes have significant implications for the accuracy and stability of measurements, which can subsequently impact load forecasting. Consequently, additional data pre-processing techniques are required to address these issues effectively.

III. Transformer Preliminary

In the context of time series analysis, the transformer architecture has been tailored to handle time series data as sequential data. Specifically, in a time series transformer, the input sequence comprises historical observations, and the model is trained to forecast future values. The self-attention mechanism [15] serves as a pivotal component within the transformer. This mechanism empowers the model to capture interdependencies among various time steps in the sequence. The transformer can effectively model long-term dependencies and capture complex patterns in the time series data by attending to relevant time steps and learning their importance.

A. Encoder and Decoder

The transformer architecture follows an encoder-decoder structure. The encoder is responsible for processing the input time series, performing high-level feature extraction, and capturing temporal dependencies. It is composed of a stack of encoder blocks, each comprising a multi-head attention module and a position-wise feed-forward network (FFN). The multi-head attention module allows the encoder to attend to different parts of the input sequence simultaneously, facilitating the extraction of relevant information. The position-wise FFN applies a non-linear transformation to each position independently, enhancing the model's ability to capture the intricate patterns within time series data. On the other hand, the decoder leverages the learned representations to generate predictions.

B. Attention Mechanism

Attention is a mechanism that allows the model to focus on different parts of the input sequence when making predictions or generating output. It enables the allocation of varying weights or importance to different elements in the sequence, enabling the model to attend to relevant information.

The attention mechanism comprises three essential elements: query $\mathbf{q}$, key $\mathbf{k}$, and value $\mathbf{v}$. The calculation of attention weights involves two main steps: 1) computing the similarity scores between the query and each element in the sequence, and 2) applying a softmax function to normalize the scores into a probability distribution. The resulting attention weights indicate how much attention or focus should be given to each input element. In the transformer architecture, the dot-product attention formulation is commonly used and can be expressed as follows:

$$Att(\mathbf{q}, \mathbf{k}, \mathbf{v}) = Softmax(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_k}})\mathbf{v}, \tag{1}$$

where $\mathbf{q}$, $\mathbf{k}$, and $\mathbf{v}$ are typically obtained by transforming the original input, and d_k represents the dimensions of keys. The attention calculation is performed for each time step within the input sequence, enabling the model to capture dependencies and patterns across different time intervals.

C. Limitations

Although the vanilla transformer-based model has been applied to time series data, there are still several limitations in our current application scenario.

First, the design of the vanilla transformer primarily emphasizes capturing local patterns and short-term dependencies, which may not adequately address the requirements of HVAC load forecasting. To achieve accurate HVAC load forecasting, it is essential to comprehend long-term global dependencies and intricate relationships between time steps. Second, the fixed-length attention mechanism limits the model's capacity to capture long-term dependencies and complex patterns that extend beyond the fixed attention window in time series load forecasting. Third, the normalization of attention mechanism weights in the input layer can result in the loss of absolute magnitude. This normalization limitation has implications for accurately representing individual time steps and can hinder the model's capability to effectively capture peak values.

These limitations highlight the need for alternative approaches that address these issues and improve the performance and applicability of HVAC load forecasting.

IV. HALO Framework Design

Building upon previous measurements and observations detailed in Sec. II, as well as the limitations we identified in the vanilla transformer outlined in Sec. III, we introduce HALO, which is a comprehensive transformer-based framework designed specifically for HVAC load forecasting. It systematically addresses not only the challenges posed by real-world HVAC data but also the limitations of the vanilla transformer.

HALO, as shown in Figure 5, consists of three main stages: an adaptive data pre-processing stage, a multi-encoders transformer-based load forecasting stage, and a scale fusion stage.

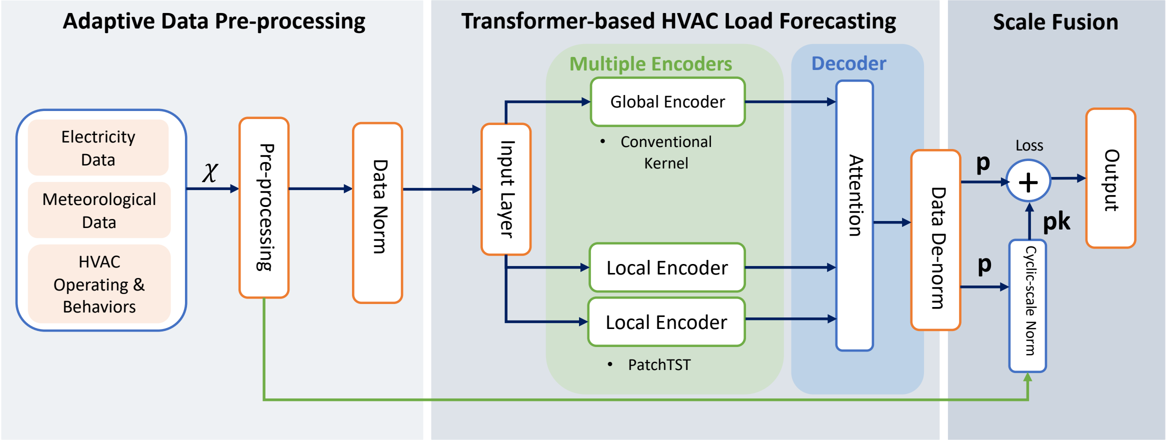


Fig. 5: HALO: Transformer-based HVAC Load Framework

The first stage is responsible for pre-processing the historical time series data, which is subsequently inputted into the second stage for accurate HVAC power load prediction. In the initial data pre-processing stage, we address data inconsistency across six buildings and mitigate high noise levels by smoothing load fluctuations caused by common electricity phenomena. The detailed design is illustrated in Section IV-A.

In the subsequent stage, we employ multiple transformer encoders and a single decoder structure for model training and load forecasting. Specifically, two types of encoders , called *global encoder* and *local encoder*, simultaneously process the input data, enhancing the effectiveness of load forecasting. The global encoder $\mathbf{z}_i^{global}$, which is implemented by SGConv [26], captures long-term dependencies and broader trends while maintaining low complexity for long-range historical data. Let j denote the index of a local encoder, we employ PatchTST [27] as backbone models for the multiple local encoders $\mathbf{z}_{i,j}^{local}$, which could efficiently capture fine-grained recent information, enabling HALO to analyze localized patterns and variations beyond the fixed attention window of the vanilla transformer model. We introduce the design in Section IV-B.

In the scale fusion stage, by employing a scale fusion strategy and corresponding loss function, HALO can enhance its generalization capabilities. This methodology effectively captures comprehensive information from the original series while emphasizing the accuracy of peak value series forecasting, achieving a better balance between them. This stage will be investigated in Section IV-C.

In the following, we present the details of the key components in the local-global-scale transformer architecture.

A. Adaptive Data Pre-processing

In the first stage illustrated in Figure 5, we conduct preliminary data processing before feeding the data into the subsequent stage. Pre-processed datasets, including electricity data, meteorological data [2], and HVAC operating and behavioral parameters. The collected data is organized in a time series format and serves as the input for training and learning in subsequent stages.

Feature Variability Handling: As discussed in Sec. II, inconsistencies due to the variability associated with different geographic locations, their environmental factors, and distinct characteristics that influence load forecasting. Specifically, buildings may have varying numbers of HVAC internal units, all controlled by an equal number of HVAC communication panels. Each HVAC communication panel collects the 37th-44th features, as specified in Table II. To handle this data variability, we aggregate the values of each feature from individual internal units.

Smoothing: In Sec. II, we have identified two common phenomena observed in HVAC systems: zero drift and voltage spikes. These phenomena introduce transient periods and load fluctuations that present challenges for accurate load predictions using existing models. To address these fluctuations, we utilize moving smoothing [28]. This approach involves replacing erratic data points with a smoother representation that effectively captures the underlying trend or pattern in the load data. By mitigating the impact of these fluctuations, we aim to improve the accuracy of load predictions, particularly during transient periods and load fluctuations caused by zero drift and voltage spikes.

B. Transformer-based HVAC Load Forecasting

In the second stage, as shown in Figure 5, we use a transformer-based encoder to capture the future global and local trends.

Model inputs and outputs Let $\mathcal{X} = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_N\}$ denote our time series dataset with length N, where each feature input $\mathbf{d}_t$ at time slot t is an M-dimensional feature vector, all features are shown in Table II. In our HVAC load forecasting scenario, let $\mathcal{P} = \{p_1, p_2, ..., p_N\}$ denotes power load values, where $p_t \in d_t$. Given an input window size L, the input $\mathbf{x}_{i:i+L-1} = \{\mathbf{d}_i, \mathbf{d}_{i+1}, ..., \mathbf{d}_{i+L-1}\}$ is fed into the model, the objective is to predict T future values $\hat{\mathbf{p}}_{i+L:i+L+T-1} = \{p_{i+L}, p_{i+L+1}, ..., p_{i+L+T-1}\}$. With the input window size L and a specific moving step S, the data can be split into $G = \lfloor \frac{N-L-T}{S} \rfloor$ groups of instance, and we have $i \in [1, G]$.

[2] The historical meteorological data was collected from https://www.visualcrossing.com/

Encoder Existing transformer-based solutions in HVAC load forecasting inadequately address the complexities inherent to real-world scenarios, such as data variability, power load fluctuations, and electronic phenomena. These solutions rely on the vanilla transformer, which has limitations like fixed-length attention, high computational complexity, and limited sequential dependency. As a result, the vanilla transformer struggles with long-term dependencies and complex patterns involving peak values. In response to these challenges, HALO incorporates a global encoder and multiple local encoders with different neural network structures and configurations. In this section, we provide a detailed explanation of these encoders and their contributions.

1) Global Convolutional Encoder: The accumulation of historical data poses the challenges in managing larger model sizes and increasing time complexity, but the vanilla transformer struggles to handle effectively. To address this, we incorporate a convolutional kernel, a specially designed matrix with sublinear complexity. This kernel extracts the global feature, enabling efficient and effective processing of extensive and noisy input signals in a long-term period [26]. A typical global convolution kernel is defined as:

$$y = u * k \tag{2}$$

where $*$ is the convolution operator, $u \in \mathbb{R}^{n \times d}$ is input sequence, $k \in \mathbb{R}^{n \times d}$ is a learnable global kernel, and $y \in \mathbb{R}^{n \times d}$ is the output.

To enable the efficient convolution kernel, we employ SGConv kernel [29] as the global encoder with multi-scale sub-kernels to capture long-range dependency from the input data more effectively and a weighted combination of sub-kernels where weights decay for larger scales. As shown in Algorithm.1, we tackle it by feeding the input sequence $\mathbf{x}_{i:i+L-1}$ into the global encoder, denoted as

$$\mathbf{z}_i^{global} = \text{GCONV}(\mathbf{x}_{i:i+L-1}) \tag{3}$$

Algorithm 1 Global Convolutional Encoder $\text{GCONV}(u)$

```
1: $L = |u|$
2: $B = [\log_2(\frac{L}{d})] + 1$
3: for i = 0 to B-1 do
4:     $k_i = \alpha^i \text{UPSAMPLE}_{2^{max[i-1,0]}d}(\mathbf{w}_i)$
5: end for
6: $k = \frac{1}{Z}[k_0, k_1, ..., k_{B-1}]$, where $Z = \text{NORM}(k)$
7: $k_f = \text{FFT}(k)$
8: $u_f = \text{FFT}(u)$
9: $y_f = \text{CONTRACT}(u_f, k_f)$
10: $y = \text{FFT}(y_f)$
11: $y = \text{NORM}(y)$
12: return $y$
```

In Algorithm.1, let L represent the length of the input sequence $\mathbf{x}_i$, the parameter set of a single channel is defined as $S = \{\mathbf{w}_i | 0 \le i < [\log_2(\frac{L}{d}]\} + 1$, where $\mathbf{w}_i \in \mathbb{R}^d$ denotes the parameter for the i-th sub-kernel k_i. The number of scales is denoted as $B = [\log_2(\frac{L}{d})] + 1$. The upsample operation is implemented using linear interpolation to create sub-kernels of different scales. $Upsample_l(\mathbf{x})$ is used to represent the upsampling of $\mathbf{x}$ to a length of l, as shown in lines 3-5. A normalization constant Z is employed to ensure that the convolution operation does not alter the scale of the input, along with a coefficient α to control the decaying speed. Then, a weighted combination scheme can be achieved by concatenating a set of weighted sub-kernels k_i, as shown in line 6. For implementation, we compute the depth-wise convolution kernel and employ Fast Fourier Transform (FFT) to perform the convolution in $O(L \log L)$ time.

2) Local Transformer Encoders: As discussed in Sec. II, we have identified challenges related to power load fluctuations in response to specific features at different time intervals. The fixed-length attention of the vanilla transformer restricts its ability to capture fluctuation trends across varying time intervals while maintaining low complexity.

To address these challenges, we propose the use of multiple local encoders designed to capture recent local information $\{\mathbf{z}_{i,j}^{local} | 1 \le j \le C\}$ at different time windows, where C denotes the number of local encoders and j is the index of the local encoder. In the j-th local encoder, we denote the patch length as h_j and the stride between two consecutive patches as s_j. Thus, let l_j denotes the number of patches, which is given by $l_j = \lfloor \frac{L-h_j}{s_j} \rfloor$.

The local encoders focus on capturing different sizes of patch length to capture dependencies as follows,

$$\mathbf{z}_{i,j}^{local} = \text{ENCODER}(\mathbf{x}_{i:i+L-1}, h_j, s_j) \tag{4}$$

where PatchTST [27] is utilized as the encoder to capture recent local information $\mathbf{z}_{i,j}^{local}$. The dimensions of $\mathbf{z}_{\text{global}}$ and $\mathbf{z}_{i,j}^{local}$ in Eq.3 and Eq.4 have a size of $[Batch \times L \times M]$, where $Batch$ is the batch size used in the experiments.

Decoder In order to leverage both global and local information effectively, our approach involves integrating the global and local information into the decoder module, which is responsible for generating prediction outcomes. This decoder module prominently includes a cross-attention module that focuses on capturing historical information within the time series. We perform a mapping of global and local features to a hidden dimension at the token level. Subsequently, we integrate the following global and local information within the decoder of a transformer model using a gating mechanism.

This allows for the effective integration of global and local information through querying the global information with the local information. The combined output is defined as follows:

$$\begin{gathered} \beta * \sum_{j=1}^{C} w_j \cdot Att_j(\mathbf{z}_i^{global}, \mathbf{z}_{i,j}^{local}, \mathbf{z}_{i,j}^{local}) + \\ (1-\beta) * \sum_{j=1}^{C} w_j \cdot Att_j(\mathbf{z}_{i,j}^{local}, \mathbf{z}_i^{global}, \mathbf{z}_i^{global}) \end{gathered} \tag{5}$$

where β is a bias between global and local encoders, w_j is the weights to balance the importance of different local encoders, $1 \le j \le C$.

The transformer decoder can effectively integrate both global and local information, allowing for a more comprehensive representation of the input sequence and potentially improving the model's ability to generate accurate predictions.

C. Scale Fusion

As observed in Section II, capturing the scale of HVAC power load, i.e., peak values, presents challenges posed by multiple factors. First, the variations in power load arise from geographic disparities, diverse HVAC system brands and types, and varying counts of internal units. Second, HVAC load is susceptible to noise due to common phenomena like zero drifts and voltage spikes. Third, power load exhibits fluctuations in response to different features within distinct time windows.

These factors make it difficult to effectively capture scale values in HVAC load forecasting, especially when normalization techniques struggle to handle the specific information embedded in cyclic correlations. This highlights the need for an innovative method tailored specifically for peak forecasting.

To address these challenges and capture the peak values, we propose the integration of a scale component. The peak values extracted from the pre-processing stage are fed into the scale fusion stage. During the scale fusion stage, the output of the transformer model, denoted as $\mathbf{p}$, is used as an input for a shallow neural network. The resulting output, denoted as $\mathbf{pk}$, is then fed into the loss function. Then we utilize a hybrid loss function to map the historical series to the original and peak value series, with a focus on the mapping relationship in the scale fusion stage. We start with a series of power load values denoted as $\mathbf{p}_{i:i+L-1} = \{p_i, p_{i+1}, ..., p_{i+L-1}\}$, where L represents the input window size. During the data pre-processing stage, we extract peak values denoted as $\mathbf{pk}_{i:i+L-1} = \{pk_1^i, ..., pk^i_{\lfloor \frac{L}{T} \rfloor}\}$ from the power load values $\mathbf{p}_{i:i+L-1}$. The peak value pk_r^i, $r \in [1, \lfloor \frac{L}{T} \rfloor]$, is obtained by downsampling in the r-th period from the sequence of $\mathbf{p}_{i:i+L-1}$ with a fixed interval T, resulting in a length of $\lfloor \frac{L}{T} \rfloor$, which is calculated by:

$$pk_r^i = max\{p_{i+(r-1)T}, ..., p_{i+rT-1}\} \tag{6}$$

The objective is to forecast the power load peak value in the next interval T. Our scale fusion strategy employs a simple yet highly effective optimization strategy [30], which simultaneously optimizes the loss function of the original time series and its corresponding peak value series.

Loss Function

We choose to utilize the Mean Squared Error (MSE) loss to quantify the difference between the predicted values and the ground truth. Consequently, we optimize the overall objective loss through the following hybrid loss function:

$$\begin{aligned} \mathcal{L} = \gamma \cdot \mathbb{E}_{\mathbf{p}} ||\hat{\mathbf{p}}_{i:i+L-1} - \mathbf{p}_{i:i+L-1}||_2^2 + \\ (1-\gamma) \cdot \mathbb{E}_{\mathbf{pk}} ||\hat{\mathbf{pk}}_{i:i+L-1} - \mathbf{pk}_{i:i+L-1}||_2^2, \end{aligned} \tag{7}$$

$\mathbb{E}\mathbf{p}$ represents the MSE loss between the ground truth original time series $\mathbf{p}_{i:i+L-1}$ and the output $\hat{\mathbf{p}}_{i:i+L-1}$ of the transformer forecasting model. On the other hand, $\mathbb{E}\mathbf{pk}$ denotes the MSE loss between the ground truth power load peak values $\mathbf{pk}_{i:i+L-1}$ and the final output $\hat{\mathbf{pk}}_{i:i+L-1}$ of the scale fusion. The weighting factor γ takes values between 0 and 1.

By employing this scale fusion strategy and corresponding loss function, HALO can enhance its generalization capabilities. This methodology allows for capturing the original series information while emphasizing the forecasting performance for the peak load values.

V. Performance Evaluation

In this section, we present the evaluation of our HALO framework with the following research questions.

RQ1: How do the local-global-scale architecture impact on the performance of HALO?

RQ2: How does the design of data pre-processing and scale fusion contribute to the performance of HALO?

RQ3: How does HALO perform when compared with state-of-the-art transformer based time series forecasting models?

Data Description. In order to evaluate the performance of the HALO framework, we employ real-world, minute-level data from EdgeSpot, an edge computing-based IoT equipment. This data encompasses public historical weather information, electricity-related metrics, and specific operational details of HVAC systems across six buildings with diverse locations and user behaviors, as shown in Table I. Missing data is addressed through the utilization of mean imputation, while outliers are handled by the smoothing step during the pre-processing stage. To meet to storage limitations and measurement requirements, we collect data at a 15-minute granularity in the experiments. This sampling interval aligns with the collection settings of the SCADA system commonly used in the power industry. We also employ the sliding window with 15 minutes to generate the data used in our experiments continuously. All data are partitioned into a training set, validation set, and test set in a ratio of 7:1:2.

Experiment Setup. In the experiment, we employed a robust server featuring dual Intel Xeon Gold 6348 CPU, 256 GB memory, and 2 NVIDIA A800 GPU. Our model training employs the ADAM optimizer [31]. We utilize a dynamic learning rate that ranges between $1e^{-4}$ and $1e^{-3}$. For training, we utilized 100 epochs. Batch size is set to 128. The hyperparameters are as follows: the weight β of balancing the importance between global encoder and local encoders is set to 0.5; in the global encoder, the decay coefficient α is chosen to be 0.5 [29], the remaining hyperparameters are the same as SGConv [26]; in the local encoder, we use two encoders to capture the recent power load information in two types of periods with the weight settings $w_1 = 0.5$ and $w_2 = 0.5$, the other hyperparameters are same to PatchTST [27]; in the scale fusion component, the weight γ in the hybrid loss function is set to 0.2. These hyperparameters were chosen based on the best performance of the proposed framework in the experiments.

Evaluation Metrics. In order to evaluate the predictive performance of the HALO model, we employed two primary indicators: the mean absolute error (MAE) and the root mean square error (RMSE) [32]. The selection of these indicators was based on their ability to accurately evaluate the model's accuracy in predicting time series data. The MAE is well-suited for time series prediction because it provides a clear, scale-sensitive measure of prediction error in the original units of the data. On the other hand, the RMSE metric measures the deviations between predicted and actual values, providing a reliable measure of predictive accuracy. The calculation methods for MAE and RMSE are presented in Eq. 8 and Eq. 9.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{9}$$

where $\hat{y}_i$ represent the predicted value of the i-th sample in the testing dataset, y_i denote the true value of the i-th sample in the test set, and n represent the total number of samples.

A. Local-Global-Scale Architecture (RQ1)

We now evaluate the effectiveness of local-global-scale architecture under various settings. As discussed in Sec. II, we noted that the power load exhibited distinct variations in response to changes in specific features across diverse time windows. This behavior served as our primary motivation to conduct an analysis of local encoders with varying patching lengths. Our aim was to ascertain their ability to capture and adapt to information from diverse time windows. In this subsection, we extend our comparison to include an investigation into the impact of using different input time window sizes of all encoders within the HALO framework.

Local Encoder Patch Length: In our experiment, we employ two local encoders with different patch lengths. One local branch uses a patch length of 60 minutes, while the other local branch uses a patch length of 720 minutes, as shown in Table III as (60;720).

Input Window Size: Our observation in Sec. II reveals that weather impacts have a noticeable effect on power load over extended periods, such as a day or week, while user behaviors primarily influence power load within shorter time windows. To capture a comprehensive range of information, we select a 1-day window (1440 minutes) and a 1-week window (10080 minutes) as input window sizes for all encoders in our experiment.

We also compare the performance of HVAC load forecasting at different time intervals to fit grid management and energy utilization in practice. We consider the following three types of predictions:

1) **Short-term Prediction**: 15-minute and 1-hour forecasts are important for balancing supply and demand, aiding in immediate operational decisions, especially in the collaboration between power load and renewable energy sources (e.g., solar and wind);

2) **Mid-term Prediction**: 12-hour and 24-hour load forecasts assist in operational planning, including the scheduling of power plants that require longer start-up times and the management of daily operations to meet anticipated demand;

3) **Long-term Prediction**: 1-week prediction enables strategic decision-making regarding energy imports and exports, long-term carbon emission reduction strategy, and renewable energy resource management over extended periods.

It is worth noting that the performance for 1-week power load prediction using a single day's data is not available, given the insufficient length of the input time series.

TABLE III: Comparison between different local branch patch sizes and input time window sizes w.r.t. MAE and RMSE.

Forecasting Time Period	Metric	Local Branch Patch Length (Minutes) (60;720)		(60;1440)	
		Input Window Size (Minutes)			
		1440	10080	1440	10080
15 Minutes	MAE	3.9637	2.2739	3.7035	**2.0397**
	RMSE	4.6262	4.2849	4.4324	**4.0692**
1 Hour	MAE	4.4393	3.9746	4.0179	**3.4803**
	RMSE	5.2479	4.5924	4.8933	**4.3498**
12 Hours	MAE	4.5839	4.2087	4.3884	**4.1528**
	RMSE	5.8365	4.6528	5.3838	**4.4893**
24 Hours	MAE	5.7475	4.4739	4.5232	**4.2566**
	RMSE	8.9468	6.0851	5.9321	**5.8367**
1 Week	MAE	-	6.3890	-	**6.2735**
	RMSE	-	7.8265	-	**7.3862**

Our findings, as presented in Table III, indicate significant improvements in MAE and RMSE, with the use of longer patch lengths in the local encoders. Specifically, using longer patch lengths resulted in respective MAE and RMSE improvements of 21.30% and 33.70% for the 24-hour forecasting time period when using a 1440-minute input window size. We also observed a similar trend in all scenarios that the performance improved as the input time window sizes increased. Specifically, compared to a 1440-minute window, a 10080-minute input window improved the MAE and RMSE by 15.44% and 12.49%, respectively for 12-hour forecasts, using a patch size of (60;1440). These findings suggest that extending the local encoders' patch lengths and increasing the input time window sizes could effectively optimize the performance of the HALO framework. This further validates our observation in Sec. II, as certain specific factors, such as temperature fluctuations, show a greater impact on the power load over longer periods. The model can capture a more comprehensive range of information by adopting larger patch lengths and expanded input time window sizes.

We also noted superior prediction performance on shorter forecasting periods. This could be attributed to the fact that longer forecasting periods require more training samples of extended input time series. Given the limitations of our data, we have fewer training samples for longer input time series, which potentially impacts the forecast performance for these periods.

B. Data-Preprocessing and Scale Fusion (RQ2)

We next evaluate the effectiveness of data pre-processing and scale fusion using an ablation study.

Data Pre-processing: Smoothing techniques play a crucial role in mitigating the impact of load fluctuations caused by zero drift and voltage spikes in real-world HVAC load forecasting. As shown in Figure 6, we present the evaluation of predictive performance using unsmoothed and smoothed

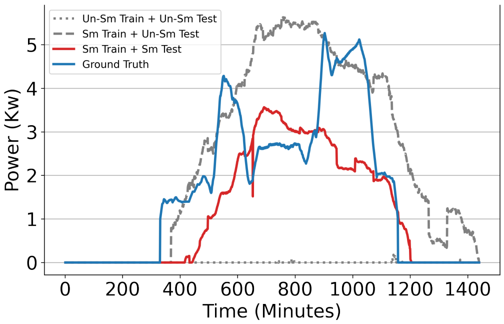


Fig. 6: Comparison of predictive performance using unsmoothed (Un-Sm) and smoothed (Sm) data for training and testing.

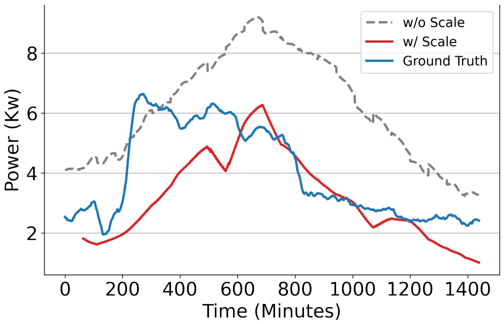


Fig. 7: Comparison of performance with (w/) and without (w/o) scale component.

data during the training and testing phases. We observed significantly inferior performance when using unsmoothed data during both the training and testing phases, likely due to the noise from load fluctuations caused by zero drift and voltage spikes. The performance significantly improved when we switched to smoothed data for both phases. Specifically, we noted improvements of 16 times and 12 times in MAE and RMSE, respectively, clearly surpassing scenarios where unsmoothed data was used. Moreover, the scenario where smoothed data was used for training and unsmoothed data was used for testing also exhibited superior performance. We observed performance enhancements of 55.86% and 44.72% on MAE and RMSE, respectively. This highlights the significant performance improvements of the data pre-processing.

We also noted superior prediction performance on shorter forecasting periods. This could be attributed to the fact that longer forecasting periods require more training samples of extended input time series. Given the limitations of our data, we have fewer training samples for longer input time series, which potentially impacts the forecast performance for these periods.

Scale Fusion: We investigate the impact of scale fusion, as seen in Figure 7. This highlights the effectiveness of the scale component to capture maximum power values in HVAC load forecasting. The MAE and RMSE of HALO with and without the scale component are 1.9015, 2.4024, and 3.7188, 4.5533, respectively. This demonstrates that the inclusion of the scale component in the HALO model leads to notable performance enhancements, yielding improvements of 95.57% and 89.53% in terms of MAE and RMSE, respectively.

C. Performance Results of HALO (RQ3)

We evaluate the performance of our proposed framework, HALO, by comparing it with the following:

Autoformer [33] is designed for time series forecasting tasks based on the transformer architecture [15]. It integrates an auto-correlation mechanism to effectively capture both long-term and short-term dependencies in time series data.

Fedformer [34] is a frequency-enhanced transformer, devised specifically for long-term prediction, leveraging the propensity of most time series data to acquire a sparse representation in well-established bases, such as the Fourier transform. It is recognized as one of the most successful transformer model variants applied to time series data [26, 34].

PatchTST [27], a channel-independent patch time series transformer, is particularly designed for multivariate time series forecasting. It borrows from the transformer model's architecture by applying patching to split the input time series into several patches. These patches are then processed by the transformer, allowing it to handle long sequences more efficiently and effectively.

Due to the substantial computational resource requirements, with each epoch taking 3-4 hours, our experiment could not be conducted using the vanilla transformer, whereas our proposed framework completes an epoch in 40-50 minutes.

To ensure a comprehensive evaluation, our experiments were also conducted across six buildings with distinct geographical locations and user behavior patterns. In this subsection, we employ the optimal local encoder patch size and input window size, as determined in subsection V-A, to forecast the 24-hour HVAC power load for effective daily operational planning and scheduling.

As shown in Table IV, our proposed framework, HALO, significantly outperforms Autoformer, Fedformer, and PatchTST, yielding average improvements of 12 times, 10 times, and 3 times, respectively across six buildings. This reinforces the model's capability to be applied more efficiently and effectively in real-world HVAC load forecasting scenarios.

TABLE IV: Comparison with state-of-the-art transformer-based time series forecasting models across six buildings for predicting 24-hour power load w.r.t. MAE and RMSE.

Buildings	Metric	HALO	Autoformer[33]	Fedformer[34]	PatchTST[27]
A	MAE	**4.2566**	149.8564	98.7653	7.8915
	RMSE	**5.8367**	197.3883	160.3028	8.5985
B	MAE	**14.7864**	30.9731	27.0892	14.8137
	RMSE	**13.7547**	41.6660	35.0941	16.9635
C	MAE	**18.1154**	48.9063	47.6588	18.2800
	RMSE	**20.0469**	68.0290	64.8167	21.1644
D	MAE	**6.0812**	29.4175	25.5368	13.5549
	RMSE	**7.7217**	38.7094	41.0608	20.8935
E	MAE	**1.0198**	32.1430	28.4530	13.2361
	RMSE	**1.6618**	27.1145	24.4982	10.5728
F	MAE	**1.9015**	21.6516	27.0497	17.2361
	RMSE	**2.4024**	28.3567	25.9818	20.5728

VI. Related Work

The relationship between reducing grid power load and carbon emissions has been empirically established, with evidence demonstrating that a decrease in power load results in a corresponding reduction in carbon emissions [9]. For optimal control of IoT devices in an HVAC system to achieve carbon emission reduction, it is important to have reliable and timely HVAC load forecasting. Existing HVAC load forecasting methods can be broadly classified into two categories: physical-based models and data-driven models. Physical-based models rely heavily on physical principles to describe the heat transfer characteristics of buildings [11]. These models, often referred to as white-box models, are capable of capturing the actual thermal response of buildings to various influential factors, including outdoor and indoor environments. However, they demand extensive and detailed building information and the prediction accuracy of these models may vary if the underlying assumptions of the physical principles are not met consistently [12]. These physical-based models highly rely on software tools such as DOE-2 [35], Designer's Simulation Toolkit (DeST) [36], etc. to simulate the energy consumption of buildings. However, acquiring proficiency in these tools typically demands a significant investment of time and effort [12]. More recently, with the continuous improvement in building energy management and the rapid advancements in artificial intelligence, a large number of black-box data-driven models based on machine learning have emerged, offering distinct advantages for building energy load forecasting [10].

A. Data-driven HVAC Load Forecasting

Over the past few decades, various shallow machine learning methods, including support vector machines (SVM) [37], multilayer perceptron (MLP) [38, 39], gradient boosting [40, 41], have been extensively investigated for HVAC load forecasting. More recently, with their advantages of adaptability, non-linearity, and the ability to handle large datasets, neural networks have gained significant popularity in HVAC load forecasting [32]. Within the realm of neural networks, convolutional neural network (CNN) [42], recurrent neural network (RNN) [43], and long short-term memory (LSTM) neural network [44] have been demonstrated to be effective in predicting the load of building HVAC systems.

Compared to individual models, integrating neural networks into hybrid prediction models has demonstrated enhanced accuracy and efficiency. Abdou *et al.* [45] discovered that the particle swarm optimization-artificial neural network (PSO-ANN) model exhibited superior performance for both heating and cooling load forecasting across three distinct climates in Morocco. Li *et al.* [46] developed a reliable building energy consumption prediction model by incorporating a genetic algorithm-neural network (GA-NN) approach. Liu *et al.* [32] proposed a hybrid model, random forest-improved sparrow search algorithm-long short-term memory (RF-ISSA-LSTM), which achieved improved accuracy and reduced running time for forecasting the cooling load of large public buildings. Guo *et al.* [41] constructed four hybrid models to enhance the prediction accuracy of heating and cooling loads.

B. Transformer-based HVAC Load Forecasting

Another type of neural network architecture, transformer, was introduced by [15] and built upon the core principle of the self-attention mechanism, utilizes matrix multiplication in its specific implementation to capture dependencies between vectors in the input sequence, regardless of their distance. The self-attention mechanism offers several advantages, including lower complexity, fewer parameters, and reduced computing power requirements. Additionally, the results of each step are independent of the results of previous steps, resulting in improved effectiveness. Transformer has demonstrated remarkable success in various Natural Language Processing (NLP) applications, such as machine translation [47] and image identification [48]. Given the shared sequential nature between time series data and NLP, the transformer has been progressively adopted for time series forecasting tasks [14, 49], demonstrating its immense potential as a reliable approach for constructing thermal load forecasting models.

Several studies have leveraged attention mechanisms to enhance prediction models within the load forecasting domain. For instance, Li *et al.* [50] proposed a novel neural network architecture with an attention mechanism for RNN-based building energy prediction, resulting in improved accuracy and interpretability in predicting the cooling load of buildings 24 hours in advance. Lim *et al.* [49] introduced a novel attention-based architecture that significantly improved performance across multiple domains, including power load forecasting. Moreover, Jurasovic *et al.* [13] and Long *et al.* [14] employed transformer-based models for load forecasting. Jurasovic *et al.* [13] presented a transformer-based load forecasting system that incorporated recent advancements in neural attention mechanisms, achieving accurate day-ahead load prediction. Li *et al.* [14] proposed a building load prediction model based on a transformer network, aiming to enhance the accuracy of building load prediction by effectively incorporating temporal dependencies. Similarly, Chen *et al.* [51] developed a transformer-based model specifically for forecasting cooling loads in airport terminals. These transformer-based models, primarily utilizing basic transformer backbones. Our proposed framework systematically address the real-world HVAC load forecasting complexities, such as data variability, power load fluctuations, and electronic phenomena.

VII. Conclusion

This paper is motivated by our initial data measurements and observations from real-world data gathered from EdgeSpot, an edge computing-based IoT device that interfaces with various IoT devices within HVAC systems. We proposed HALO, a transformer-based framework for HVAC load forecasting, to address the challenges observed in real-world data and existing data-driven methods. HALO integrates an adaptive data pre-processing stage, a transformer-based load forecasting stage with multiple encoders, and a scale fusion stage. This integration enables the accurate prediction of HVAC load, which in turn optimizes energy efficiency in HVAC operations and leads to a reduction in carbon emissions [9]. Through extensive evaluation using real-world data traces from a prototype application, we demonstrated that HALO consistently performs well across multiple buildings, thereby contributing to the overarching goal of reducing carbon emissions. In our future work, our plan is to leverage the outcomes of our HVAC load forecasting to formulate intelligent control strategies. By employing EdgeSpot, we aim to regulate a variety of IoT devices to promote a sustainable and efficient energy future.

VIII. Acknowledgement

The research was supported in part by an RGC RIF grant under the contract R6021-20, RGC CRF grants under the contracts C7004-22G and C1029-22G, and RGC GRF grants under contracts 16209120, 16200221, 16207922, 17203320 and 17209822.

References

[1] "For a livable climate: Net-zero commitments must be backed by credible action," https://www.un.org/en/climatechange/net-zero-coalition, accessed: 2023-08-11.

[2] "Emissions by sector," https://www.iea.org/energy-system/buildings, accessed: 2023-08-10.

[3] D. Biswas, "Reinforcement learning based hvac optimization in factories," in *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*, 2020, pp. 428–433.

[4] D. Minoli, K. Sohraby, and B. Occhiogrosso, "Iot considerations, requirements, and architectures for smart buildings—energy optimization and next-generation building management systems," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 269–283, 2017.

[5] X. Zhang, M. Pipattanasomporn, T. Chen, and S. Rahman, "An iot-based thermal model learning framework for smart buildings," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 518–527, 2019.

[6] M. Avci, M. Erkoc, A. Rahmani, and S. Asfour, "Model predictive hvac load control in buildings using real-time electricity pricing," *Energy and Buildings*, vol. 60, pp. 199–209, 2013.

[7] G. Lowry, "Day-ahead forecasting of grid carbon intensity in support of heating, ventilation and air-conditioning plant demand response decision-making to reduce carbon emissions," *Building Services Engineering Research and Technology*, vol. 39, no. 6, pp. 749–760, 2018.

[8] F. Qian, W. Gao, Y. Yang, and D. Yu, "Potential analysis of the transfer learning model in short and medium-term forecasting of building hvac energy consumption," *Energy*, vol. 193, p. 116724, 2020.

[9] "Greenhouse gases equivalencies calculator - calculations and references," https://www.epa.gov/energy/greenhouse-gases-equivalencies-calculator-calculations-and-references, accessed: 2024-02-10.

[10] Z. Gao, J. Yu, A. Zhao, Q. Hu, and S. Yang, "A hybrid method of cooling load forecasting for large commercial building based on extreme learning machine," *Energy*, vol. 238, p. 122073, 2022.

[11] C. Fan, F. Xiao, and Y. Zhao, "A short-term building cooling load prediction method using deep learning algorithms," *Applied energy*, vol. 195, pp. 222–233, 2017.

[12] Y. Huang and C. Li, "Accurate heating, ventilation and air conditioning system load prediction for residential buildings using improved ant colony optimization and wavelet neural network," *Journal of Building Engineering*, vol. 35, p. 101972, 2021.

[13] M. Jurasovic, E. Franklin, M. Negnevitsky, and P. Scott, "Day ahead load forecasting for the modern distribution network-a tasmanian case study," in *2018 Australasian Universities Power Engineering Conference (AUPEC)*. IEEE, 2018, pp. 1–6.

[14] L. Long, S. Xingyu, X. Bi, L. Yueliang, and X. Sun, "A novel transformer-based network forecasting method for building cooling loads," *Energy and Buildings*, p. 113409, 2023.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] A. Teke and O. Timur, "Assessing the energy efficiency improvement potentials of hvac systems considering economic and environmental aspects at the hospitals," *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 224–235, 2014.

[17] H. Sun, S. Li, F. R. Yu, Q. Qi, J. Wang, and J. Liao, "Toward communication-efficient federated learning in the internet of things with edge computing," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11 053–11 067, 2020.

[18] M. V. Ngo, T. Luo, and T. Q. Quek, "Adaptive anomaly detection for internet of things in hierarchical edge computing: A contextual-bandit approach," *ACM Transactions on Internet of Things*, vol. 3, no. 1, pp. 1–23, 2021.

[19] R. Li, Q. Li, J. Zhou, and Y. Jiang, "Adriot: an edge-assisted anomaly detection framework against iot-based network attacks," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 10 576–10 587, 2021.

[20] H. Siraj, S. M. Shaharyar, and N. Arshad, "Economic dispatch incorporating the greenness index of energy generation sources," in *Proceedings of the Ninth International Conference on Future Energy Systems*, 2018, pp. 441–443.

[21] R. Jha, S. Lee, S. Iyengar, M. H. Hajiesmaili, D. Irwin, and P. Shenoy, "Emission-aware energy storage scheduling for a greener grid," in *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*, 2020, pp. 363–373.

[22] "Weather data api," https://www.visualcrossing.com/, accessed: 2023-09-12.

[23] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Springer Berlin Heidelberg, 2009, pp. 1–4.

[24] D. Li, Y. Wang, J. Wang, C. Wang, and Y. Duan, "Recent advances in sensor fault diagnosis: A review," *Sensors and Actuators A: Physical*, vol. 309, p. 111990, 2020.

[25] D. O. Johnson and K. A. Hassan, "Issues of power quality in electrical systems," *International Journal of Energy and Power Engineering*, vol. 5, no. 4, pp. 148–154, 2016.

[26] Y. Zhao, Z. Ma, T. Zhou, L. Sun, M. Ye, and Y. Qian, "Gcformer: An efficient framework for accurate and scalable long-term multivariate time series forecasting," *arXiv preprint arXiv:2306.08325*, 2023.

[27] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.

[28] R. J. Hyndman, "Moving averages," in *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg, 2011, pp. 866–869.

[29] Y. Li, T. Cai, Y. Zhang, D. Chen, and D. Dey, "What makes convolutional models great on long sequence modeling?" *arXiv preprint arXiv:2210.09298*, 2022.

[30] Z. Zhang, X. Wang, J. Xie, H. Zhang, and Y. Gu, "Unlocking the potential of deep learning in peak-hour series forecasting," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 4415–4419.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] Z. Liu, J. Yu, C. Feng, Y. Su, J. Dai, and Y. Chen, "A hybrid forecasting method for cooling load in large public buildings based on improved long short term memory," *Journal of Building Engineering*, vol. 76, p. 107238, 2023.

[33] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 419–22 430, 2021.

[34] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *International Conference on Machine Learning*. PMLR, 2022, pp. 27 268–27 286.

[35] D. A. York and C. C. Cappiello, "Doe-2 engineers manual (version 2. 1a)," Lawrence Berkeley Lab., CA (USA); Los Alamos National Lab., NM (USA), Tech. Rep., 1981.

[36] X. Zhang, J. Xia, Z. Jiang, J. Huang, R. Qin, Y. Zhang, Y. Liu, and Y. Jiang, "Dest—an integrated building simulation toolkit part ii: Applications," in *Building Simulation*, vol. 1. Springer, 2008, pp. 193–209.

[37] R. E. Edwards, J. New, and L. E. Parker, "Predicting future hourly residential electrical consumption: A machine learning case study," *Energy and Buildings*, vol. 49, pp. 591–603, 2012.

[38] J. Massana, C. Pous, L. Burgas, J. Melendez, and J. Colomer, "Short-term load forecasting in a non-residential building contrasting models and attributes," *Energy and Buildings*, vol. 92, pp. 322–330, 2015.

[39] Z. Yan, X. Zhu, X. Wang, Z. Ye, F. Guo, L. Xie, and G. Zhang, "A multi-energy load prediction of a building using the multi-layer perceptron neural network method with different optimization algorithms," *Energy Exploration & Exploitation*, vol. 41, no. 1, pp. 273–305, 2023.

[40] Z. Wang, T. Hong, and M. A. Piette, "Building thermal load prediction through shallow machine learning and deep learning," *Applied Energy*, vol. 263, p. 114683, 2020.

[41] J. Guo, S. Yun, Y. Meng, N. He, D. Ye, Z. Zhao, L. Jia, and L. Yang, "Prediction of heating and cooling loads based on light gradient boosting machine algorithms," *Building and Environment*, vol. 236, p. 110252, 2023.

[42] H. Sun, Y. Niu, C. Li, C. Zhou, W. Zhai, Z. Chen, H. Wu, and L. Niu, "Energy consumption optimization of building air conditioning system via combining the parallel temporal convolutional neural network and adaptive opposition-learning chimp algorithm," *Energy*, vol. 259, p. 125029, 2022.

[43] I. Metsä-Eerola, J. Pulkkinen, O. Niemitalo, and O. Koskela, "On hourly forecasting heating energy consumption of hvac with recurrent neural networks," *Energies*, vol. 15, no. 14, p. 5084, 2022.

[44] Y. Li, Z. Tong, S. Tong, and D. Westerdahl, "A data-driven interval forecasting model for building energy prediction using attention-based lstm and fuzzy information granulation," *Sustainable Cities and Society*, vol. 76, p. 103481, 2022.

[45] N. Abdou, Y. El Mghouchi, K. Jraida, S. Hamdaoui, A. Hajou, and M. Mouqallid, "Prediction and optimization of heating and cooling loads for low energy buildings in morocco: An application of hybrid machine learning methods," *Journal of Building Engineering*, vol. 61, p. 105332, 2022.

[46] X. Li, S. Liu, L. Zhao, X. Meng, and Y. Fang, "An integrated building energy performance evaluation method: From parametric modeling to ga-nn based energy consumption prediction modeling," *Journal of Building Engineering*, vol. 45, p. 103571, 2022.

[47] Y. Li, J. Li, and M. Zhang, "Deep transformer modeling via grouping skip connection for neural machine translation," *Knowledge-Based Systems*, vol. 234, p. 107556, 2021.

[48] Z.-M. Chen, Q. Cui, B. Zhao, R. Song, X. Zhang, and O. Yoshie, "Sst: Spatial and semantic transformers for multi-label image recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 2570–2583, 2022.

[49] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[50] A. Li, F. Xiao, C. Zhang, and C. Fan, "Attention-based interpretable neural network for building cooling load prediction," *Applied Energy*, vol. 299, p. 117238, 2021.

[51] B. Chen, W. Yang, B. Yan, and K. Zhang, "An advanced airport terminal cooling load forecasting model integrating ssa and cnn-transformer," *Energy and Buildings*, p. 114000, 2024.

Cheng Pan received her B.Comm. degree in Management Information Systems from the University of Alberta in 2016, and her MPhil in Computer Science from the University of Hong Kong in 2023. Between 2016 and 2021, she worked as a data specialist in the healthcare industry in Canada. She is currently pursuing her Ph.D. at the University of Hong Kong, supervised by Dr. Edith Ngai. Her research interests include the Internet of Things, artificial intelligence, and machine learning.

Cong Zhang received M.Sc. from Zhengzhou University, Zhengzhou, China, in 2012, and Ph.D. from Simon Fraser University, Canada, in 2018. His research interests include deep learning for smart energy, multimedia communications, and cloud/edge computing.

Edith C.H. Ngai is an Associate Professor in the Department of Electrical and Electronic Engineering at the University of Hong Kong. Before joining HKU in 2020, she was an Associate Professor in the Department of Information Technology at Uppsala University, Sweden. Her research interests include Internet-of-Things, edge intelligence, smart cities, and smart health. She was a VINNMER Fellow (2009) awarded by the Swedish Governmental Research Funding Agency VINNOVA. Her co-authored papers received a Best Paper Award in QShine 2023 and Best Paper Runner-Up Awards in IEEE IWQoS 2010 and ACM/IEEE IPSN 2013. She was an Area Editor of IEEE Internet of Things Journal from 2020 to 2022. She is currently an Associate Editor of Ad Hoc Networks, Computer Networks, IEEE Transactions of Mobile Computing, and IEEE Transactions of Industrial Informatics. She served as a program chair in ACM womENcourage 2015 and a TPC co-chair in IEEE SmartCity 2015, IEEE ISSNIP 2015, and IEEE GreenCom 2022. She received a Meta Policy Research Award in Asia Pacific in 2022. She was one of the N^{2}Women Stars in Computer Networking and Communications in 2022. She is a Distinguished Lecturer in the IEEE Communication Society in 2023-2024.

Jiangchuan Liu (S'01-M'03-SM'08-F'17) is a University Professor in the School of Computing Science, Simon Fraser University, British Columbia, Canada. He is a Fellow of The Canadian Academy of Engineering, an IEEE Fellow, and an NSERC E.W.R. Steacie Memorial Fellow. He was an EMC-Endowed Visiting Chair Professor at Tsinghua University (2013-2016). In the past, he worked as an Assistant Professor at The Chinese University of Hong Kong and as a research fellow at Microsoft Research Asia. He received the BEng degree (cum laude) from Tsinghua University, Beijing, China, in 1999, and the PhD degree from The Hong Kong University of Science and Technology in 2003, both in computer science. He is a co-recipient of the inaugural Test of Time Paper Award of IEEE INFOCOM (2015), ACM SIGMM TOMCCAP Nicolas D. Georganas Best Paper Award (2013), and ACM Multimedia Best Paper Award (2012). His research interests include multimedia systems and networks, cloud and edge computing, social networking, online gaming, and Internet of Things/RFID/backscatter. He has served on the editorial boards of IEEE/ACM Transactions on Networking, IEEE Transactions on Big Data, IEEE Transactions on Multimedia, IEEE Communications Surveys and Tutorials, and IEEE Internet of Things Journal. He is a Steering Committee member of IEEE Transactions on Mobile Computing and Steering Committee Chair of IEEE/ACM IWQoS (2015-2017). He is TPC Co-Chair of IEEE INFOCOM'2021.

Bo Li is a Chair Professor in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. He was a Cheung Kong Visiting Chair Professor in Shanghai Jiao Tong University (2010-2016), and was the Chief Technical Advisor for ChinaCache Corp. (NASDAQ:CCIH), a leading CDN service provider. He made pioneering contributions in multimedia communications and the Internet video broadcast, in particular the Coolstreaming system, which was credited as the first large-scale Peer-to-Peer live video streaming system in the world. It attracted significant attention from both industry and academia with over USD 25M investment and received the Test-of-Time Best Paper Award from IEEE INFOCOM (2015). He received 6 Best Paper Awards from IEEE including IEEE INFOCOM (2021). He has been an editor or a guest editor for over a two dozen of IEEE and ACM journals and magazines. He was the Co-TPC Chair for IEEE INFOCOM 2004.

He is a Fellow of IEEE. He received his PhD in the Electrical and Computer Engineering from University of Massachusetts at Amherst, and his B. Eng. (summa cum laude) in the Computer Science from Tsinghua University, Beijing, China.