Embedding Periodic Patterns into Tokens: Masked Autoencoder for Household Electricity Demand Forecasting

Yuugou Ohno[1,2], **Tomonori Honda**[1], **Masaki Onishi**[1], **Norihiro Itsubo**[2]

[1]National Institute of Advanced Industrial Science and Technology (AIST)

[2]Waseda University

yuugou.oono@aist.go.jp, tomonori-honda@aist.go.jp,masaki@aist.go.jp, itsubo-n@waseda.jp

Abstract

To support the global initiative on energy efficiency, developing highly accurate household-level electricity demand forecasting models is crucial. Various methods, including statistical approaches and models based on LSTMs and Transformers, have been explored for this purpose. However, these models often face challenges due to distribution shifts from the complex overlapping periodicities inherent in the electricity demand data. To address this issue, we propose an autoencoder that integrates periodic time-series information into each token, effectively capturing the intricate characteristics of electricity demand data. The method transforms time-series data into a two-dimensional representation by segmenting it according to its representative periodicities. The segmented data is then divided into patches, linearly embedded, randomly masked, and reconstructed at the point level. Experimental results on household electricity demand datasets demonstrate that our approach, which embeds periodic information at the token level, significantly enhances forecasting performance.

1. Introduction

In Japan's 2016 Global Warming Countermeasures Plan, which the Cabinet approved, the residential sector was assigned the highest emission reduction target among all industries, with an objective of a 66% reduction from 2013 levels (Government of Japan 2021). This highlights the urgency of reducing greenhouse gas emissions associated with residential energy consumption. In parallel, significant progress is being made in establishing an electricity trading market, demand response (DR) mechanisms, and a power capacity market. Accurate electricity demand forecasting is essential to effectively leverage these energy markets for efficient energy control. As illustrated in Figure 1, the original data (top left) exhibits complex noise resulting from the interplay of multiple overlapping periodic patterns. These patterns include an annual seasonal cycle (bottom left), a weekly cycle distinguishing weekdays and weekends (top right), and a daily 24-hour cycle (bottom left).

Recently, Transformer-based approaches have demonstrated considerable success in time series forecasting tasks.

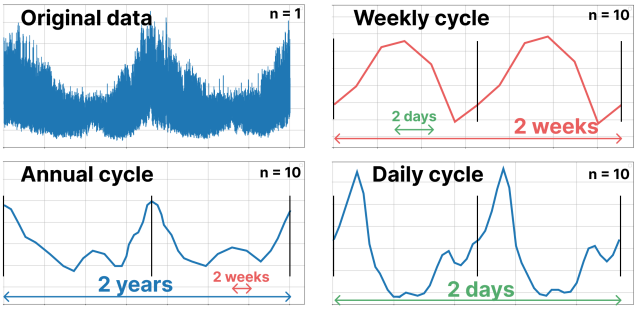

Figure 1: Multiple periodicity in electricity demand and smoothing of data by increasing the number of households

Current Transformer models analyze patterns from past segments of time series data and extrapolate these patterns to predict future values(Zhou et al. 2021). However, these generative Transformer models primarily rely on masking and reconstructing future values. These approaches are relatively inefficient, limiting their learning capacity and introducing significant distribution shift issues. Several models have attempted to address these challenges by decomposing the input time series into trend and seasonal components. However, many non-stationary time series, such as electricity demand, exhibit complex cyclic patterns that necessitate intricate, hierarchical preprocessing steps, which are computationally expensive. Figure 1 illustrates that the power consumption data comprises of multiple cycles patterns.

Zhe Li and colleagues effectively addressed the challenges mentioned above in 2023 (Li et al. 2023). The proposed model utilizes a larger portion of the input data, thereby mitigating distribution shift issues by masking specific sections of the time series during the training phase. This model randomly masks portions of the time series data during training and employs an autoencoder to reconstruct these masked sections at the individual point level. Ti-MAE has demonstrated high predictive accuracy across several publicly available real-world datasets, including electricity demand data.

Electricity demand data is highly complex, exhibiting multiple intertwined periodicities. Previous research has established that transforming such data into a two-dimensional format aligned with its periodic cycles can enhance the rep-

resentation of these complex features. This study proposes a novel approach that maximizes the potential of Ti-MAE for electricity demand forecasting. By converting time series data into a two-dimensional format, it captures multiple periodic cyclic patterns, enabling the model to learn and predict from these intricate temporal structures effectively.

2. Related work

2.1 electricity demand forecasting

Rodrigues et al. conducted a systematic review of modeling methods for short-term electricity demand forecasting within the residential sector (Rodrigues et al. 2023). They gathered 334 relevant studies from four major databases—Web of Science, IEEE Xplore, Scopus, and Science Direct—spanning a decade (2012–2022) and thoroughly analyzed 38 English-language studies using the PRISMA methodology.

In studies on residential electricity demand, various approaches are employed contingent upon the unit of analysis—individual households, single multi-residential buildings, or groups of such residential units. Research findings indicate that forecasting electricity demand at the residential cluster level yields higher accuracy than prediction based on individual households. This enhanced accuracy results from the aggregation of consumption patterns, which smooths out fluctuations observed at the level of individual households.

The two graphs on the left side of Figure 1 represent two years of electricity demand data: one for a single household and the other for an average across 1,000 households. The graph labeled "Original data" shows hourly electricity demand forecasts for a single household over two years, revealing a high level of noise that obscures clear patterns. In contrast, the graph labeled "Annual cycle", derived by averaging the data of 1,000 households, demonstrates a smoother trend, allowing key features to be more easily identified.

Additionally, increasing the number of residential clusters has been shown to reduce the standard error of cascade predictions. Conversely, predicting electricity demand at the individual household level remains a considerable challenge, largely due to the complexity of capturing characteristics unique to each household. To enhance scalability, households are typically separated between training and test datasets, exacerbating distribution shift issues.

2.2 Ti-MAE

Li et al. proposed Ti-MAE as a novel framework to improve the prediction accuracy of time series data (Li et al. 2023). Similar to other autoencoders, Ti-MAE comprises an encoder that maps the observed time series signal $X \in \mathbb{R}^{T \times m}$ to a latent representation $H \in \mathbb{R}^{T \times n}$, and a decoder that reconstructs the original sequence from the embeddings generated by the encoder at each timestamp (Figure 2). Ti-MAE employs an asymmetric design where the encoder applies masking to the input embeddings and processes only the visible tokens. A lighter decoder then operates on the encoded tokens augmented with masked tokens to reconstruct the original time series at the point level. This framework was evaluated across five real-world datasets, predicting time series values 12, 24, 48, 96, 128, and 168 hours ahead. The results indicated superior performance compared with representation learning methods and other Transformer-based approaches.

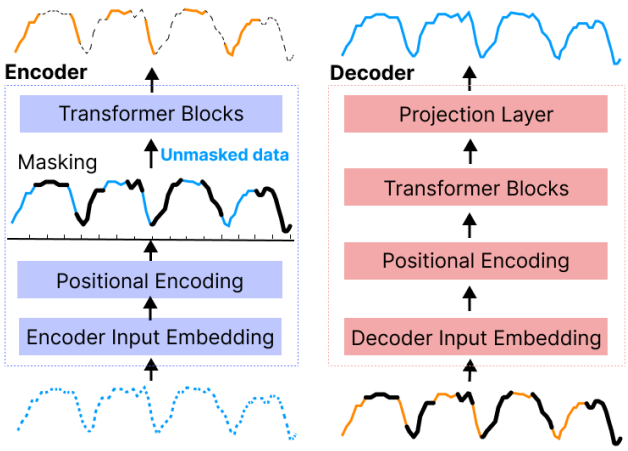

Figure 2: Ti-MAE structure overview

3. Methodology

3.1 Problem definition

Let $X = (x_1, x_2, \ldots, x_T) \in \mathbb{R}^{T \times m}$ represent the electricity demand data for m households over a length of T, where each m corresponds to an individual household. Given a historical univariate time series segment $X_h \in \mathbb{R}^h$ of length h for a single household, the forecasting task aims to predict the subsequent k step values, denoted as $X_f \in \mathbb{R}^k$.

3.2 Model architecture

Figure 3 provides an overview of the model structure. The architecture closely resembles that of Ti-MAE, with the key distinction being the pre-processing of time series data into a two-dimensional format before encoding. This transformation enables each token to represent information across multiple cyclical periods.

In this structure, the two-dimensional time series data is segmented into patches, and then embedded through linear projection. Positional encoding is subsequently applied to the embedded patches, followed by masking, with only unmasked tokens processed by the encoder. The decoder reconstructs the original time series at the level of individual points after it has received the encoded visible patches and masked tokens. Unlike other MAE-based models, this model introduces an asymmetric design to optimize reconstruction. A detailed description of each component is given in the following sections.

Transformation to Two-Dimensional Data The discrete Fourier transform (DFT) is used to calculate the frequency components of the time series data to obtain the maximum wavelength W. Dividing the time series data into segments of W units and arranging these segments into H columns

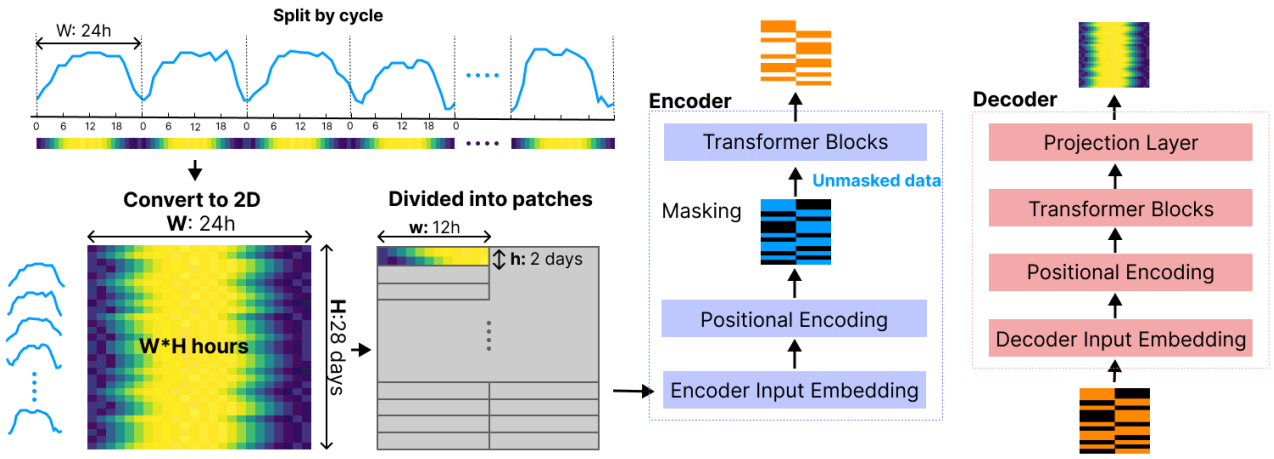

Figure 3: Proposed method structure overview

transforms into a two-dimensional format with dimensions $W \times H$. This transformation is performed mechanically, without requiring prior knowledge of the data or manual intervention. Consequently, we can consider two distinct dimensions in terms of W and H.

Patch Embedding The two-dimensional data is divided into patch size $w \times h$, with each patch treated as a single token. This approach enables each patch to encompass time-series information across both the W and the H directions. Under the default setting, w is set to 12 and h to 2, corresponding to data preceding the past 12 hours and the same 12-hour period from the previous day.

To preserve positional information, each patch is embedded using a linear projection augmented with fixed sinusoidal positional embeddings. By avoiding the manual embedding of dates or data-specific values, this approach minimizes the introduction of functional biases.

Masking After tokenizing the resulting $\frac{W}{w} \times \frac{H}{h}$ patches, a subset of tokens is randomly sampled without replacement, while the remaining tokens are masked. The optimal masking ratio depends largely on the information density and redundancy of the data and significantly affects the model's performance. Generally, image data exhibits greater spatial redundancy than natural language, as individual pixels often lack significant meaning. This characteristic is similarly observed in time series data. When handling data with low information content, applying a higher masking ratio helps prevent the model from overemphasizing low-level semantic content. For instance, BERT, a widely used model in natural language processing, uses a masking rate of 15%, while image generation models and Ti-MAE employ a 75% masking ratio. In our experiments, a 75% masking ratio was found to be optimal.

Encoder The encoding procedure employs a standard Vision Transformer (ViT) to encode only the visible tokens, following embedding and random masking (Dosovitskiy 2020). This approach is similar to other Masked Autoencoder (MAE) designs with asymmetric architectures, resulting in a significant reduction in both time complexity and memory usage relative to full encoding (He et al. 2022).

Decoder The decoder consists of an additional set of Transformer blocks applied to the full set of tokens, which includes both the visible tokens generated by the encoder and the masked tokens. The masked tokens are represented as shared learned vectors that indicate the patches to be predicted. The top layer of the decoder consists of a linear projection that reconstructs the input by predicting all point-level values. The loss function is the mean squared error (MSE), calculated between the original time series data and the reconstructed measurements.

4. Experiments

4.1 Experimental setup

Datasets This study utilized hourly electricity demand data derived from smart meters provided by a housing manufacturer. Specifically, the analysis employed panel data comprising 1,000 households with uninterrupted records over two years, from January 1, 2020, to December 31, 2021, corresponding to approximately 1.75×10^8 time steps. Although the dataset also includes other time-series information such as building characteristics and weather conditions, this study focused exclusively on electricity demand owing to its straightforward data collection process.

To represent the data as images, the time series of a single household was denoted as T, the resized image dimensions as $T_s \times T_l$, and the number of images as N. The transformation process involved generating images by applying a sliding window over the time series, where the top-left corner of the window corresponds to the beginning of the series and the bottom-right corner corresponds to the end. The starting point of the sliding window was randomize to ensure that the spatial position of the image did not consistently correspond to specific days or times. During the training phase, the images were divided into patches of dimensions $W \times H$, with 75% of the patches masked for prediction tasks. In the

testing phase, the minimum number of patches required to reconstruct the desired output k was generated. By default, the image dimensions were set to $T_s \times T_l = 24 \times 28$, and the patch size was set to $W \times H = 12 \times 2$ as shown in Figure 3.

The generated images were split into training, validation, and testing datasets using a 7:1:2 ratio, with the division conducted at the household level. To align with the potential real-world application, where predictions may be required for households without prior historical data, target households designated for prediction were excluded from the training data.

Baselines This study employed these baseline methods: LSTM (Guo et al. 2021), a machine learning-based approach commonly applied in this domain; and Ti-MAE (Li et al. 2023) which served as a comparative benchmark.

Implementation Details The encoder and decoder architectures each consisted of eight Transformer blocks, incorporating 16-head self-attention mechanisms. The hidden layer dimensionality was configured to 512, and the model was trained with a batch size of 256. The MSE was employed as the loss function.

During the forecasting phase, patches sufficient for predicting h-hour intervals were generated, and predictions were compared with actual values to calculate the MSE was used as an evaluation metric, given its widespread use in electricity demand forecasting. Let y denote the actual consumption at a given time, $\hat{y}$ the predicted value, and n the total sample size.

All the models are implemented using PyTorch and trained and tested on a single NVIDIA A100-SXM4-40GB GPU.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

4.2 Result of Time Series Forcasting

Comparison of Forecasting Accuracy Across Methods Table 1 illustrates that the proposed method consistently reduces the forecasting error across all prediction horizons, ranging from 12 hours to 168 hours. Notably, the method achieves a maximum MSE reduction of 8.1% compared with Ti-MAE and 64.71% compared with LSTM. Figure 4 presents the MSE values for each forecasting method across various time horizons. The accuracy improvements over LSTM are more particularly notable for longer forecasting periods. While LSTM exhibits a substantial decline in accuracy as the prediction horizon increases, the proposed method demonstrates robustness, with only a slight decrease in performance.

Furthermore, the proposed method and Ti-MAE can use a single-trained model to predict multiple horizons. In contrast, LSTM require separate models for each horizon, increasing complexity. The proposed method demonstrates improved accuracy over Ti-MAE across all forecasting horizons. This improvement is attributed to the inclusion of time-series positional information in the token representations of the proposed method. Other parameters, such as input size and window size, remain identical. These results indicate that incorporating positional information into tokens enhances forecasting accuracy.

Table 1: Accuracy comparison between the proposed method and baseline methods

Time Metric	Proposed	Ti-MAE	LSTM
12h	**0.1121**	0.1164	0.1578
24h	**0.1153**	0.1193	0.1650
48h	**0.1116**	0.1204	0.1827
168h	**0.1199**	0.1252	0.3403

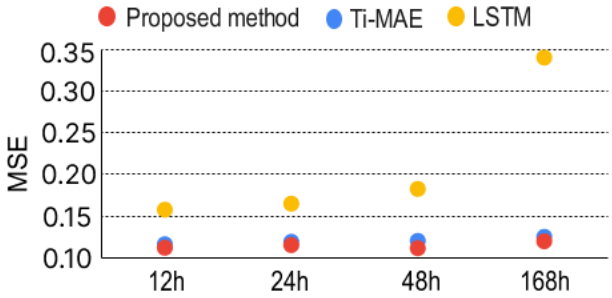

Figure 4: MSE of forecasting for the proposed method and baseline methods

The Effect of Providing Periodic Information to Tokens The following section presents a detailed comparison between Ti-MAE and the proposed method. Table 2 and Figure 5 present experimental results obtained by varying the method of incorporating information for cases where tokens represent 12-hours and 24-hours data, respectively. The notation 12(2,6) indicates that each token represents a total of 12 hours of information, comprising a consecutive 6-hour segment and the corresponding 6-hour segment from the previous day. Similarly, the notation 24(2,12) denotes that each token represents 24 hours of information, consisting of a continuous 12-hour segment and the corresponding 12-hour segment from the previous day.

In contrast, the Ti-MAE partitions the consecutive time series into patches and tokenizes them, indicating that it includes data from a consecutive 12-hour period, represented as 12(1,12).

Figure 5 presents the MSE values for each forecasting method across different time horizons. Although the total amount of information per token remains constant, changes in the integration method resulted in up to a 14.6% reduction in MSE. Across all forecasting horizons, methods that include time series information from one day prior (12(2,6) or 24(2,12)) achieved higher accuracy than those that relied solely on consecutive time series, such as Ti-MAE (12(1,12) or 24(1,24)).

In Ti-MAE, increasing the token's information from 12

Table 2: Accuracy comparison of proposed method and Ti-MAE for different configurations

Time	Proposed Method		Ti-MAE	
Metric	12 (2, 6)	24 (2, 12)	12 (1, 12)	24 (1, 24)
12h	**0.1096**	**0.1121**	0.1236	0.1164
24h	**0.1145**	**0.1153**	0.1266	0.1193
48h	**0.1111**	**0.1116**	0.1277	0.1204
168h	**0.1190**	**0.1199**	0.1373	0.1252

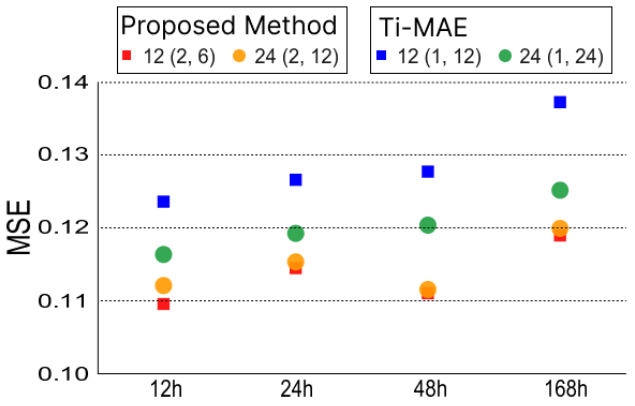

Figure 5: MSE of forecasting for each token size and encoded information

hours to 24 hours improved accuracy. In contrast, the proposed method achieves nearly identical accuracy with 12(2,6) and 24(2,12). Experiments with 12(4,3) and 24(4,6), which include additional daily information, revealed a decrease in accuracy. These results indicate the importance of striking a balance between periodic information and continuous time series data.

However, these results indicate that modifying the temporal positions represented within the tokens impacts accuracy. Incorporating multiple temporal periods at the token level allows the model enables capture features of the time series data that are not fully discernible through the Transformer's attention mechanisms alone.

Accuracy Comparison with Varying Input Times Experiments were conducted by varying the input length of the model. Figure 6 plots the average MSE for predictions spanning 12 to 168 hours across different input lengths. The highest accuracy was achieved with an input length of 672 hours (corresponding to an image of size 28×24). The shortest input length of 384 hours (16×24) resulted in significantly lower accuracy. For other input lengths, the accuracy showed little variation and remained stable. These results indicate that for short-term predictions of approximately 168 hours a moderate input length is sufficient to achieve reliable accuracy.

The optimal masking ratio The optimal masking ratio was found to be 0.75 (75%), aligning with the findings from Ti-MAE. This ratio is comparable to those used for images.

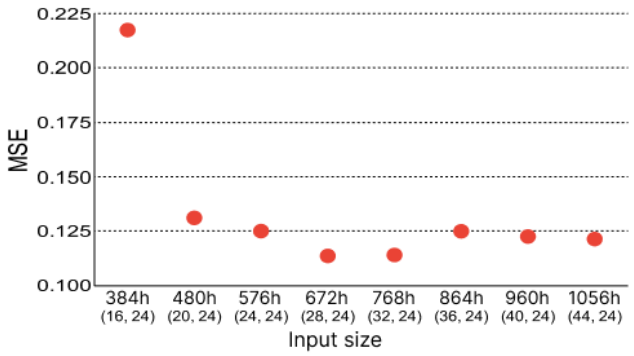

Figure 6: Average MSE of forecasting for four-time horizons (12 h, 24 h, 48 h, 168 h) across different input sizes

A high masking ratio allows the model can efficiently learn features from time series data, which typically contain less semantic information.

5. Conclusion

In this study, electricity demand forecasting was conceptualized as a multi-decadal seasonality task. This study proposed a novel framework: a masked autoencoder that embeds periodicity as a feature of token representation. This model enables the effective handling of multiple periodicities, even in the presence of distribution shifts.

Experimental results showed that the proposed method outperformed existing approaches over extended prediction horizons compared with existing methods. Specifically, it achieved a 14.6% reduction in MSE compared to other known methods, including Ti-MAE and LSTM. The proposed method effectively captured intricate temporal patterns by embedding a range of periodicities at the token level. The technique also proved robust and flexible, retaining good performance over varying input and output lengths.

However, the methodology has limitations. This study focused exclusively on univariate electrical demand forecasting, centered on the electricity demand variable, without incorporating extraneous factors such as weather conditions or household characteristics. Future studies could explore incorporating these variables to improve forecasting accuracy. Despite this limitation, the methodology is designed to address broader applications in real-world scenarios across multiple time series datasets. Future investigations could explore its applicability to domains such as weather forecasting and financial market analysis, both of which exhibit periodicity in their time series.

This study contributes new techniques to time series forecasting practice and sets the stage for more accurate and efficient modeling of complex periodic data.

Acknowledgement

We appreciate Ichijo Co., Ltd. for providing the data. This paper is based on results obtained from a project, JPNP14004, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Government of Japan. 2021. Plan for Global Warming Countermeasures. Cabinet Decision on October 22, 2021.

Guo, X.; Gao, Y.; Li, Y.; Zheng, D.; and Shan, D. 2021. Short-term household load forecasting based on Long- and Short-term Time-series network. *Energy Reports*, 7: 58–64. ICPE 2020-The International Conference on Power Engineering.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Li, Z.; Rao, Z.; Pan, L.; Wang, P.; and Xu, Z. 2023. Ti-mae: Self-supervised masked time series autoencoders. *arXiv preprint arXiv:2301.08871*.

Rodrigues, F.; Cardeira, C.; Calado, J. M. F.; and Melicio, R. 2023. Short-Term Load Forecasting of Electricity Demand for the Residential Sector Based on Modelling Techniques: A Systematic Review. *Energies*, 16(10).

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.