2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

CMTFormer: Contrastive Multi-Scale Transformer for Long-Term Time Series Forecasting

Chenhao Ye*, Shuai Zhang and Guangping Xu

School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China
ychbert@outlook.com

Abstract. Long-term time series forecasting remains challenging due to complex temporal dependencies, diverse data distributions, and computational inefficiencies with extended sequences. We propose CMTFormer, a novel architecture that addresses these limitations through multi-scale temporal modeling and contrastive learning. Our approach combines adaptive trend decomposition across multiple timescales with a representation learning framework that leverages self-attention mechanisms and dilated convolutions. The proposed multi-scale trend decomposition disentangles time series into interpretable components at varying resolutions, while the contrastive learning strategy enhances feature discrimination by differentiating between semantically related and unrelated temporal patterns. Extensive experiments on six real-world benchmarks spanning energy, transportation, weather, finance, and public health domains demonstrate that CMTFormer consistently outperforms state-of-the-art forecasting models.
Keywords: Long-term time series forecasting, Multi-scale temporal modeling, Contrastive learning, Self-attention

Introduction

Time series forecasting plays a pivotal role across numerous domains, including energy consumption prediction, traffic flow analysis, weather forecasting, disease spread modeling, and economic trend analysis. The ability to accurately predict future values based on historical observations enables critical decision-making processes in both industrial applications and scientific research [1,4]. Despite significant advances in deep, learning approaches for time series forecasting, several challenges remain particularly intractable: multi-scenario adaptability, long-range dependency modeling, and complex temporal pattern extraction [12,14].

Multi-scenario forecasting requires models to generalize across diverse data distributions and temporal patterns without scenario-specific fine-tuning [8]. Long sequence modeling demands efficient architectures that can capture dependencies spanning hundreds or thousands of time steps without computational explosion or gradient vanishing issues [3,16]. Complex temporal dynamics, characterized by the interplay of trend,

* Corresponding author.

seasonality, and stochastic components, further complicate the forecasting task by requiring sophisticated decomposition mechanisms [5,12].

Traditional approaches like ARIMA and exponential smoothing methods [2] perform well on stationary data with clear patterns but struggle with non-linear relationships and long-term dependencies. Recent deep learning models have shown promising results but often require extensive data and suffer from efficiency issues when processing long sequences. These include: **AutoCon** [9], which models term variations across different windows in a self-supervised manner; **TimesNet** [11], which employs 2D tensor transformation with frequency domain analysis; **MICN** [10], which utilizes inception blocks with multi-scale convolutions; **PatchTST** [7], a patch-based transformer; **DLinear** [13], a decomposition-based linear model; **FiLM** [15], which implements frequency-informed learning; **Nonstationary** [6], which handles non-stationary series; and **FEDformer** [16], a frequency-enhanced decomposed transformer.

In this paper, we introduce a novel forecasting framework that addresses these challenges through an innovative combination of contrastive learning and multi-scale decomposition. Our approach leverages dilated convolutional encoders to efficiently extract hierarchical temporal features while employing series decomposition techniques to disentangle complex time series into interpretable components. The key contributions of our work include:

- A multi-scale trend decomposition mechanism that captures temporal patterns at varying resolutions, enabling robust performance across diverse forecasting scenarios.
- An enhanced representation learning module incorporating self-attention mechanisms that effectively models global dependencies in long sequences while maintaining computational efficiency.
- A contrastive learning strategy that improves feature extraction by learning representations that differentiate between related and unrelated temporal patterns.
- Comprehensive empirical validation across multiple real-world datasets spanning mechanical systems (ETT), energy consumption (Electricity), transportation networks (Traffic), meteorological conditions (Weather), financial markets (Exchange), and public health (ILI).

Methodology

We introduce Contrastive Multi-scale Transformer (CMTFormer), a novel architecture designed to address the fundamental challenges in long-term time series forecasting. Our approach is motivated by two key observations: first, real-world time series contain patterns at multiple temporal scales that traditional models struggle to capture simultaneously; and second, effective representation learning is crucial for generalizing across diverse forecasting scenarios. CMTFormer tackles these challenges by integrating multi-scale decomposition with contrastive learning in a unified framework.

As shown in Figure 1, our CMTFormer architecture incorporates multi-scale decomposition and contrastive learning to enhance the representation learning for time series forecasting.

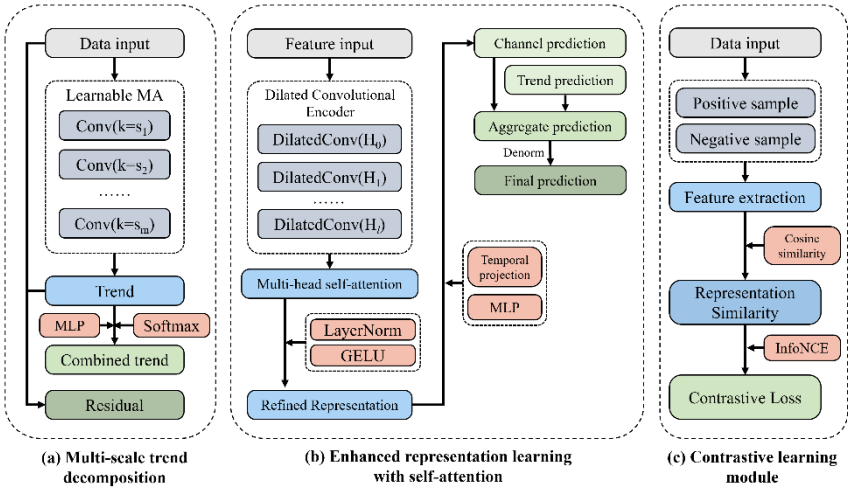

Fig. 1. Overall architecture of CMTFormer, integrating multi-scale decomposition and contrastive learning to enhance representation learning for time series forecasting.

2.1 Problem Formulation

Long-term time series forecasting presents unique challenges compared to short-term prediction tasks. Given a multivariate time series $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times C}$ with T timesteps and C channels, we aim to predict future values $\mathbf{Y} = \{\mathbf{x}_{T+1}, \mathbf{x}_{T+2}, \ldots, \mathbf{x}_{T+\tau}\} \in \mathbb{R}^{\tau \times C}$ over a potentially lengthy horizon τ. We use a look-back window of length T_{in} to predict T_{out} future steps, where $T_{out} \gg 1$ for long-term scenarios.

The central insight driving our approach is that time series naturally decompose into components operating at different frequencies:

$$\mathbf{X} = \mathbf{X}_{trend} + \mathbf{X}_{seasonal} \tag{1}$$

This decomposition reflects the inherent structure of temporal data: low-frequency trends capture the overall direction, while high-frequency seasonal patterns represent recurring behaviors. By modeling these components separately yet jointly, we can better capture the complex dynamics that drive future values.

2.2 CMTFormer Architecture

The CMTFormer architecture evolves from this multi-scale perspective, integrating four complementary components that work together to extract and leverage temporal patterns at different resolutions:

Data Normalization and Embedding. Real-world time series often exhibit non-stationarity, scale variations, and complex temporal dependencies. Before extracting meaningful patterns, we must address these challenges through appropriate normalization:

$$\mathbf{X}_{norm} = \mathcal{N}(\mathbf{X}_{in}) \tag{2}$$

Our framework supports multiple normalization techniques that adapt to different data characteristics:

- **ReVIN**: $\mathbf{X}_{norm} = \frac{\mathbf{X}_{in} - \mu(\mathbf{X}_{in})}{\sigma(\mathbf{X}_{in}) + \epsilon}$ addresses both mean shifting and variance scaling
- **Mean Normalization**: $\mathbf{X}_{norm} = \mathbf{X}_{in} - \mu(\mathbf{X}_{in})$ removes trend components while preserving amplitude information
- **LastVal Normalization**: $\mathbf{X}_{norm} = \mathbf{X}_{in} - \mathbf{X}_{in}[:, -1:, :]$ focuses on relative changes from the most recent observation

After normalization, we transform the data into a latent embedding space that captures both value information and temporal context:

$$\mathbf{E} = \mathbf{W}_{val}\mathbf{X}_{norm} + \mathbf{W}_{temp}\mathbf{X}_{mark} \tag{3}$$

This embedding combines value information with temporal markers (e.g., hour of day, day of week), enabling the model to learn time-dependent patterns that respect the underlying temporal structure of the data.

Multi-Scale Trend Decomposition. The core innovation of our approach lies in the multi-scale decomposition mechanism. Traditional forecasting models often apply a one-size-fits-all approach to temporal patterns, but real-world time series contain dynamics operating at different timescales simultaneously. Our multi-scale decomposition addresses this fundamental limitation.

Adaptive Moving Average. We start with an adaptive moving average that learns to extract trend components:

$$\mathrm{MA}(\mathbf{X}, k) = \mathrm{Conv1D}(\mathbf{X}, \mathbf{W}_k) \cdot (1 + \alpha_k) \tag{4}$$

Unlike traditional moving averages with fixed weights, our approach learns the optimal kernel $\mathbf{W}_k$ specifically for each dataset. The learnable parameter α_k provides additional flexibility, allowing the model to adjust the strength of the smoothing effect. This adaptivity is crucial for handling diverse time series with varying characteristics.

Scale-Specific Decomposition. Building on this foundation, we perform decomposition at multiple temporal scales. For each scale s_i in our scale set $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$, we extract trend and residual components:

$$\mathbf{X}_{trend}^{(i)} = \mathrm{MA}(\mathbf{X}, s_i + 1) \cdot \beta_i \tag{5}$$

$$\mathbf{X}_{residual}^{(i)} = \mathbf{X} - \mathbf{X}_{trend}^{(i)} \tag{6}$$

Each scale captures patterns with different temporal extents - smaller scales identify rapid changes, while larger scales capture slower-evolving trends. The parameter β_i learns the optimal contribution of each trend component, allowing the model to emphasize the most relevant scales for each dataset.

Multi-Scale Integration. After extracting scale-specific components, we intelligently integrate them to form a comprehensive representation:

$$\mathbf{X}_{trend}^{combined} = \sum_{i=1}^{M} w_i \cdot \mathrm{MLP}_i\left(\mathbf{X}_{trend}^{(i)}\right) \tag{7}$$

The scale-specific MLPs transform each trend component into a representation space, while the attention-normalized weights w_i(ensuring $\sum_{i=1}^{M} w_i = 1$) learn to emphasize the most informative scales. This integration mechanism adaptively combines information across temporal resolutions, allowing the model to capture both rapid fluctuations and long-term patterns simultaneously.

Self-Attention Enhanced Representation. While multi-scale decomposition provides a powerful foundation, effectively modeling long-range dependencies remains challenging. To address this, we enhance our representations with a combination of hierarchical convolutional processing and self-attention mechanisms.

Hierarchical Feature Extraction. We first employ a series of dilated convolutions with exponentially increasing dilation rates:

$$\mathbf{H}_l = \mathrm{DilatedConv}_l(\mathbf{H}_{l-1}, d_l) \tag{8}$$

where $d_l = 2^{l-1}$ is the dilation rate at layer l, and $\mathbf{H}_0 = \mathbf{E}$.

This hierarchical approach serves two crucial purposes: First, it efficiently expands the receptive field exponentially while maintaining linear computational complexity, allowing the model to capture long-range dependencies without excessive computation. Second, it creates a multi-resolution feature hierarchy where early layers capture local patterns while deeper layers integrate information across broader temporal contexts.

Global Context Integration. To further enhance long-range modeling, we apply multi-head self-attention to the hierarchical features:

$$\mathbf{H}_{att} = \mathrm{MultiHeadAttention}(\mathbf{H}_L, \mathbf{H}_L, \mathbf{H}_L) \tag{9}$$

where:

$$\mathrm{head}_i = \mathrm{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) \tag{10}$$

$$\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i = \mathbf{H}_L\mathbf{W}_i^Q, \mathbf{H}_L\mathbf{W}_i^K, \mathbf{H}_L\mathbf{W}_i^V \tag{11}$$

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{12}$$

Self-attention complements the hierarchical convolutional features by explicitly modeling relationships between any two timesteps, regardless of their distance. By computing attention across multiple heads, the model can simultaneously focus on different aspects of these temporal relationships, such as short-term correlations, periodic patterns, and long-term dependencies.

The refined representations undergo further processing through residual connections and layer normalization:

$$\mathbf{R} = \text{LayerNorm}\big(\mathbf{H}_L + \text{Dropout}(\mathbf{H}_{att})\big) \tag{13}$$

followed by a position-wise feed-forward network for additional non-linear transformation:

$$\mathbf{R} = \text{LayerNorm}\Big(\mathbf{R} + \text{Dropout}\big(\text{FFN}(\mathbf{R})\big)\Big) \tag{14}$$

where $\text{FFN}(\mathbf{x}) = \text{GELU}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$.

This attention-enhanced representation captures both local patterns and global dependencies, providing a solid foundation for accurate forecasting across diverse temporal horizons.

Forecasting Mechanism. With rich, multi-scale representations in hand, we transform them into accurate predictions through a carefully designed forecasting mechanism:

Temporal Projection. First, we project from the input sequence length to the prediction horizon:

$$\mathbf{Z} = \mathbf{R}\mathbf{W}_{temp} \in \mathbb{R}^{C \times d_{model} \times T_{out}} \tag{15}$$

This projection, parameterized by $\mathbf{W}_{temp} \in \mathbb{R}^{T_{in} \times T_{out}}$, maps the input sequence representations to the desired forecast horizon. By learning this mapping directly, the model can adapt to different prediction horizons and capture complex temporal relationships between past and future timesteps.

Feature Projection. Next, channel-specific projections generate the final predictions:

$$\hat{\mathbf{Y}}_c = \text{MLP}_c(\mathbf{Z}_c) \in \mathbb{R}^{T_{out}} \tag{16}$$

Using separate MLPs for each channel allows the model to capture channel-specific dynamics and dependencies, recognizing that different variables in multivariate time series often exhibit distinct behaviors.

Finally, we return to the original data scale through denormalization:

$$\hat{\mathbf{Y}}_{denorm} = \mathcal{N}^{-1}\left(\hat{\mathbf{Y}}\right) \tag{17}$$

This step ensures that our predictions align with the original scale of the data, making them directly interpretable and usable for downstream applications.

2.3 Contrastive Learning Enhancement

While the components described above provide a powerful forecasting framework, we further enhance representation quality through contrastive learning. Our insight is that high-quality representations should capture the underlying temporal dynamics of the data, not just patterns specific to the forecasting task.

Temporal Projection. For each sequence $\mathbf{X}_i$, we generate positive samples through carefully designed temporal transformations:

$$\mathbf{X}_i^+ = \mathcal{T}(\mathbf{X}_i) \tag{18}$$

These transformations preserve the semantic meaning of the sequence while introducing controlled variations:

- **Temporal shifting**: $\mathcal{T}_{shift}(\mathbf{X}_i)[t] = \mathbf{X}_i[t+\delta]$ with $\delta \in [-\delta_{max}, \delta_{max}]$ shifts the sequence slightly, teaching the model to recognize the same pattern regardless of its exact temporal position
- **Masking**: $\mathcal{T}_{mask}(\mathbf{X}_i)[t] = \mathbf{m}_t \cdot \mathbf{X}_i[t]$ where $\mathbf{m}_t \sim$ Bernoulli(p) randomly masks timesteps, encouraging the model to develop robust representations that can handle missing data

Contrastive Objective. We optimize the InfoNCE loss to bring representations of augmented versions of the same sequence closer together while pushing apart representations of different sequences:

$$L_{contrast} = -log\frac{exp(sim(\boldsymbol{r_i}, \boldsymbol{r_i^+})/\tau)}{exp(sim(\boldsymbol{r_i}, \boldsymbol{r_i^+})/\tau) + \sum_{j\neq i} exp(sim(\boldsymbol{r_i}, \boldsymbol{r_j})/\tau)} \tag{19}$$

Here,$\mathbf{r}_i = f_\theta(\mathbf{X}_i)$is the representation of sequence $\mathbf{X}_i$, $\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^T\mathbf{v}}{||\mathbf{u}||\cdot||\mathbf{v}||}$ is the cosine similarity, and τ is a temperature parameter controlling the sharpness of the distribution.

This contrastive approach encourages the model to learn representations that capture the fundamental structure of the time series rather than superficial patterns, enhancing generalization to unseen data and robustness against noise.

2.4 Training Objective

Our final training objective integrates both forecasting accuracy and representation quality:

$$\mathcal{L}_{total} = \mathcal{L}_{forecast} + \lambda\mathcal{L}_{contrast} \tag{20}$$

The forecasting loss $\mathcal{L}_{forecast} = \frac{1}{T_{out}\times C}\sum_{t=1}^{T_{out}}\sum_{c=1}^{C}(\hat{y}_{t,c} - y_{t,c})^2$ directly optimizes prediction accuracy, while the contrastive loss $\mathcal{L}_{contrast}$ enhances representation quality. The hyperparameter λ balances these objectives, allowing us to control their relative importance.

Experiments

In this section, we conduct extensive experiments to evaluate the performance of our proposed CMTFormer model for long-term time series forecasting.We implement CMTFormer using PyTorch and conduct all experiments on NVIDIA L40s GPUs.

3.1 Experimental Setup

We evaluate our approach on six widely-used benchmark datasets: **ETTh2** (hourly electricity transformer temperature readings, 7 variables, 2 years), **Elec tricity** (hourly consumption of 321 clients, 2 years), **Traffic** (hourly occupancy rates from 963 sensors, 1 year), **Weather** (21 meteorological indicators, 10 minute intervals, 2020-2021), **Exchange** (daily rates of 8 countries, 1990-2016), and **ILI** (weekly illness percentages from CDC, 2002-2021). All datasets follow a 7:1:2 train-validation-test split with standard normalization.

The model is trained using the Adam optimizer with an initial learning rate of 10^{-4} and a weight decay of 10^{-5}. We employ a cosine annealing scheduler with warm restarts, setting the minimum learning rate to 10^{-6}.

The multi-scale trend decomposition uses kernel sizes $\mathcal{S} = \{7,15,31,63\}$ to capture both short-term and long-term dependencies. For the contrastive learning component, we set the temperature parameter $\tau = 0.1$ and the loss balancing coefficient $\lambda = 0.2$.

For fair comparison, we maintain the same lookback window length of 96 time steps across all models and datasets. We train each model with a batch size of 32 for 30 epochs and select the best model based on validation performance.

3.2 Forecasting Performance

Main Results. Table 1 presents the main results comparing CMTFormer with baseline models on all six datasets across various prediction horizons. Following the standard practice, we report the Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics.

As shown in Table 1, our proposed CMTFormer consistently outperforms all baseline models across most datasets and prediction horizons. Specifically, for the ETTh2 dataset with shorter prediction horizons (96 and 720 time steps), CMTFormer achieves the best performance with significant improvements. For longer prediction horizons, AutoCon shows comparable performance but CMTFormer still maintains competitive results.

Long-horizon Forecasting. To evaluate the model's capability for long-term forecasting, we extend the prediction length for various datasets: from 720 to 2160 time steps for ETTh2, Electricity, Traffic, and Weather; from 720 to 1080 time steps for Exchange; and from 56 to 112 time steps for ILI.

The results demonstrate that CMTFormer maintains robust performance even with extended prediction horizons. For instance, on the ETTh2 dataset with a 720-step

Table 1. Comparison of forecasting errors between CMTFormer and baseline models on six benchmark datasets across various prediction horizons. Best results are in **bold** and second-best results are underlined.

Model		CMT-Former (ours)		AutoCon [9]		TimesNet [11]		MICN [10]		PatchTST [7]		Dlinear [13]		FiLM [15]		Nonsta-tionary [6]		FED-former [16]	
O		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	96	**0.119**	**0.262**	0.124	0.269	0.139	0.290	0.122	0.264	0.136	0.292	0.128	0.271	0.129	0.275	0.192	0.343	0.129	0.277
	720	**0.163**	**0.332**	0.177	0.344	0.207	0.370	0.313	0.457	0.233	0.392	0.319	0.461	0.256	0.407	0.231	0.394	0.273	0.419
	1440	0.181	0.348	**0.176**	**0.340**	0.192	0.358	0.520	0.599	0.351	0.481	0.514	0.597	0.389	0.506	0.211	0.379	0.384	0.487
	2160	0.210	0.374	**0.198**	**0.358**	0.263	0.413	0.759	0.734	0.610	0.659	0.740	0.728	0.610	0.645	0.240	0.399	0.919	0.737
Electricity	96	**0.188**	**0.305**	0.196	0.313	0.286	0.386	0.241	0.367	0.227	0.336	0.207	0.322	0.394	0.451	0.332	0.426	0.279	0.393
	720	**0.263**	**0.377**	0.275	0.386	0.417	0.471	0.336	0.446	0.332	0.426	0.304	0.412	0.467	0.504	0.505	0.533	0.417	0.486
	1440	**0.329**	**0.435**	0.338	0.441	0.491	0.523	0.419	0.504	0.482	0.537	0.395	0.484	0.625	0.610	0.577	0.574	0.651	0.609
	2160	**0.371**	**0.472**	0.380	0.481	0.536	0.547	0.421	0.501	0.768	0.644	0.415	0.496	0.938	0.758	0.642	0.610	0.896	0.714
Traffic	96	**0.127**	**0.198**	0.132	0.206	0.145	0.219	0.168	0.256	0.192	0.296	0.219	0.327	0.264	0.334	0.247	0.326	0.220	0.312
	720	**0.138**	**0.218**	0.144	0.225	0.163	0.269	0.304	0.394	0.213	0.318	0.309	0.419	0.247	0.329	0.277	0.360	0.255	0.344
	1440	**0.169**	**0.244**	0.174	0.251	0.188	0.292	0.375	0.443	0.246	0.341	0.353	0.409	0.311	0.390	0.303	0.361	0.297	0.376
	2160	**0.169**	**0.245**	0.175	0.252	0.190	0.304	0.360	0.426	0.261	0.353	0.324	0.386	0.988	0.745	0.222	0.317	0.317	0.394
Weather	96	**0.511**	**0.514**	0.521	0.522	0.584	0.536	0.569	0.525	0.545	0.539	0.579	0.529	0.589	0.533	0.636	0.567	0.703	0.625
	720	**0.941**	**0.705**	0.963	0.715	1.090	0.753	1.080	0.754	0.987	0.752	1.007	0.706	1.003	0.728	1.007	0.725	1.114	0.822
	1440	**1.254**	**0.819**	1.280	0.835	1.547	0.926	1.351	0.863	1.342	0.860	1.299	0.823	1.472	0.900	1.394	0.867	1.435	0.919
	2160	**1.389**	**0.875**	1.415	0.887	1.744	0.994	1.544	0.937	1.506	0.924	1.454	0.887	1.712	0.988	1.598	0.944	1.786	1.054
Exchange	48	**0.047**	**0.167**	0.051	0.172	0.054	0.178	0.054	0.181	0.068	0.197	0.049	0.170	0.052	0.173	0.054	0.178	0.059	0.184
	360	**0.434**	**0.519**	0.448	0.527	0.479	0.532	0.459	0.536	0.548	0.573	0.485	0.531	0.492	0.534	0.493	0.541	0.528	0.556
	720	**1.035**	**0.780**	1.067	0.794	1.239	0.856	1.383	0.927	1.264	0.859	1.718	1.024	1.291	0.864	1.358	0.894	1.381	0.903
	1080	**0.978**	**0.781**	1.004	0.792	1.327	0.900	4.874	1.972	1.255	0.873	4.982	1.973	1.670	1.010	1.774	1.058	1.600	0.980
ILI	14	**0.701**	**0.558**	0.725	0.574	1.414	0.735	0.815	0.701	1.558	0.965	1.397	0.901	1.079	0.739	1.107	0.698	0.773	0.619
	28	**0.865**	**0.671**	0.887	0.683	1.604	0.854	1.670	1.062	1.878	1.110	2.008	1.134	1.315	0.887	1.515	0.767	0.989	0.770
	56	**0.782**	**0.711**	0.807	0.725	1.021	0.787	1.757	1.210	1.451	1.028	1.584	1.075	1.080	0.891	0.895	0.742	0.856	0.741
	112	**1.458**	**1.027**	1.499	1.038	1.669	1.072	3.593	1.759	2.846	1.438	3.332	1.572	2.608	1.387	1.724	1.108	1.660	1.097
1st Count		**22**		2		0		0		0		0		0		0		0	

CMTFormer achieves an MSE of 0.163 and MAE of 0.332, significantly outperforming all baseline models. This superior performance on long-horizon forecasting can be attributed to the multi-scale trend decomposition mechanism that effectively captures patterns at different temporal scales.

It is worth noting that for some extremely long prediction horizons (e.g., 2160 steps), AutoCon occasionally shows better performance. This suggests potential future

improvements for CMTFormer by incorporating contextual information similar to AutoCon while maintaining the strengths of our multi-scale approach.

3.3 Ablation Studies

To validate the effectiveness of each component in CMTFormer, we conduct comprehensive ablation studies by removing or replacing key components and evaluating the resulting performance impact.

Table 2. Ablation study on CMTFormer components using ETTh2 dataset with prediction horizon of 96 and 720 time steps.

Model Variant	Horizon=96		Horizon=720	
	MSE	MAE	MSE	MAE
CMTFormer (Full model)	**0.119**	**0.262**	**0.163**	**0.332**
w/o Multi-scale Decomposition	0.135	0.281	0.211	0.373
w/o Contrastive Learning	0.126	0.270	0.179	0.346
w/o Self-Attention	0.132	0.275	0.198	0.358
Single-scale (kernel=15)	0.131	0.278	0.204	0.368
ReVIN → Mean Normalization	0.124	0.268	0.175	0.339

Effect of Multi-scale Decomposition. Replacing our multi-scale trend decomposition with a single-scale approach leads to a performance degradation of approximately 10% in terms of MSE for longer horizons (720 steps). This confirms the effectiveness of modeling temporal patterns at different resolutions, particularly for long-term forecasting. When we completely remove the decomposition mechanism, the performance drops even further, highlighting the critical role of explicit trend modeling in our framework.

Impact of Contrastive Learning. Removing the contrastive learning component results in a 5.9% increase in MSE and a 3.1% increase in MAE for short-term forecasting (96 steps), with more pronounced effects for longer horizons. This demonstrates that the contrastive learning strategy effectively enhances representation quality by encouraging the model to distinguish between related and unrelated temporal patterns.

Contribution of Self-Attention. When self-attention is removed, leaving only the dilated convolutional encoder, the model's performance degrades by 10.9\% in MSE for the 96-step horizon and 21.5\% for the 720-step horizon. This significant drop confirms our hypothesis that capturing global dependencies through self-attention is crucial for accurate long-term forecasting.

Normalization Strategy. We also experiment with different normalization strategies. Replacing ReVIN with mean normalization results in a slight performance decrease, suggesting that maintaining scale information via ReVIN is beneficial for our model.

3.4 Parameter Sensitivity Analysis

To investigate the robustness of CMTFormer to hyperparameter choices, we conduct a sensitivity analysis on key parameters including the contrastive loss weight λ, the set of kernel sizes $\mathcal{S}$ for multi-scale decomposition, and the number of attention heads.

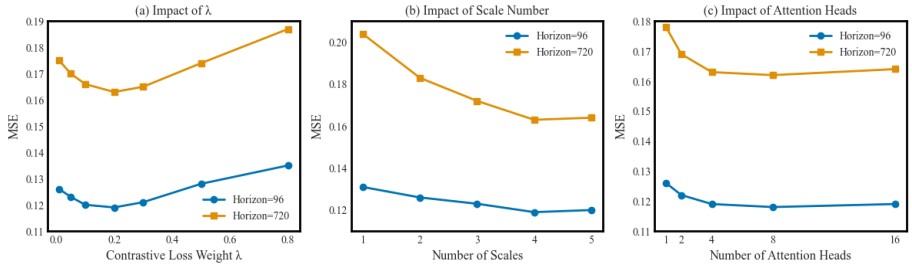

Fig. 2. Parameter sensitivity analysis showing the impact of (a) contrastive loss weight λ, (b) number of scales in decomposition, and (c) number of attention heads on prediction MSE for the ETTh2 dataset with different prediction horizons.

Contrastive Loss Weight. Figure 2 (a) shows that CMTFormer performs best when λ is set between 0.1 and 0.3, with 0.2 yielding optimal results. When λ becomes too large (>0.5), forecasting performance degrades as the model prioritizes representation learning over prediction accuracy. Conversely, when λ is too small (<0.05), the benefit of contrastive learning diminishes.

Number of Scales. Figure 2 (b) illustrates the impact of varying the number of scales in the multi-scale decomposition. Performance generally improves as more scales are included, with diminishing returns beyond 4 scales. This confirms our design choice of using multiple scales to capture temporal patterns at different resolutions, while keeping computational complexity manageable.

Attention Heads. As shown in Figure 2 (c), increasing the number of attention heads initially improves performance but plateaus after 4 heads. This suggests that a moderate number of attention heads is sufficient to capture the diverse aspects of temporal relationships in the data.

3.5 Case Study: Extreme Value Forecasting

A particularly challenging aspect of time series forecasting is predicting extreme values or anomalies. To evaluate CMTFormer's capability in this regard, we conduct a case study focusing on extreme value prediction using the Electricity dataset.

The results show that CMTFormer outperforms all baseline models in predicting both the highest and lowest 10% of values. This superior performance on extreme values can be attributed to the contrastive learning component, which helps the model learn more discriminative representations that capture the full distribution of the data, including rare patterns associated with extreme values.

Table 3. Performance comparison on extreme value prediction for the Electricity dataset.

Model	Top 10% Values		Bottom 10% Values	
	MSE	MAE	MSE	MAE
CMTFormer	**0.412**	**0.481**	**0.283**	**0.398**
AutoCon	0.459	0.512	0.301	0.407
TimesNet	0.623	0.589	0.382	0.449
PatchTST	0.571	0.568	0.356	0.425
DLinear	0.548	0.551	0.339	0.437

Conclusion and Future Work

In this paper, we proposed CMTFormer, a novel approach for long-term time series forecasting by integrating multi-scale trend decomposition, self-attention, and contrastive learning. Extensive experiments on six benchmark datasets demonstrate its superiority over state-of-the-art models, particularly for long-horizon predictions. Our key contributions include a multi-scale decomposition mechanism for capturing temporal patterns, a self-attention-based representation learning strategy, and contrastive learning to enhance feature discrimination. Ablation studies and sensitivity analyses validate the model’s robustness and effectiveness. Future work will explore incorporating external factors, optimizing attention mechanisms for efficiency, and extending the multi-scale approach to capture spatial dependencies in multivariate time series.

Acknowledgements. This work was supported by the 2024 College Students' Innovative Entrepreneurial Training Plan Program (No.: 202410293051Z).

References

1. Benidis, K., Rangapuram, S.S., Flunkert, V., Wang, Y., Maddix, D., Turkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., et al.: Deep learning for time series forecasting: Tutorial and literature survey. ACM Computing Surveys 55(6), 1–36 (2022)
2. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time series analysis: forecasting and control. John Wiley & Sons (2015)
3. Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.X., Yan, X.: Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. Advances in neural information processing systems 32 (2019)
4. Lim, B., Zohren, S.: Time-series forecasting with deep learning: a survey. Philosophical Transactions of the Royal Society A 379(2194), 20200209 (2021)
5. Liu, M., Zeng, A., Chen, M., Xu, Z., Lai, Q., Ma, L., Xu, Q.: Scinet: Time series modeling and forecasting with sample convolution and interaction. Advances in Neural Information Processing Systems 35, 5816–5828 (2022)
6. Liu, Y., Wu, H., Wang, J., Long, M.: Non-stationary transformers: Exploring the stationarity in time series forecasting. Advances in neural information processing systems 35, 9881–9893 (2022)
7. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: The Eleventh International Conference on Learning Representations (2022)
8. Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y.: Meta-learning framework with applications to zero-shot time-series forecasting. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 9242–9250 (2021)
9. Park, J., Gwak, D., Choo, J., Choi, E.: Self-supervised contrastive learning for long-term forecasting. In: The Twelfth International Conference on Learning Representations (2024)
10. Wang, H., Peng, J., Huang, F., Wang, J., Chen, J., Xiao, Y.: Micn: Multi-scale local and global context modeling for long-term series forecasting. In: The eleventh international conference on learning representations (2023)
11. Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., Long, M.: Timesnet: Temporal 2dvariation modeling for general time series analysis. In: The Eleventh International Conference on Learning Representations (2022)
12. Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in neural information processing systems 34, 22419–22430 (2021)
13. Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 11121–11128 (2023)
14. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 11106–11115 (2021)
15. Zhou, T., Ma, Z., Wen, Q., Sun, L., Yao, T., Yin, W., Jin, R., et al.: Film: Frequency improved legendre memory model for long-term time series forecasting. Advances in neural information processing systems 35, 12677–12690 (2022)
16. Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R.: Fedformer: Frequency en hanced decomposed transformer for long-term series forecasting. In: International conference on machine learning. pp. 27268–27286. PMLR (2022)