# FECAM: Frequency Enhanced Channel Attention Mechanism for Time Series Forecasting

# FECAM：频率增强通道注意力机制用于时间序列预测

Pengyu Zeng Kai Wang\*

彭宇森 汪凯\*

Shenyang Institute of Automation, Shenyang Institute of Automation,

沈阳自动化研究所 沈阳自动化研究所，

Chinese Academy of Sciences Chinese Academy of Sciences

中国科学院 中国科学院

Shenyang, China Shenyang, China

沈阳，中国 沈阳，中国

University of Chinese Academy of wangkai@sia.cn

中国科学院大学 wangkai@sia.cn

Maowei Jiang

江茂伟

Shenyang Institute of Automation,

沈阳自动化研究所，

Chinese Academy of Sciences

中国科学院

Shenyang, China

沈阳，中国

University of Chinese Academy of

中国科学院大学

Sciences

北京，中国

Beijing, China

科学

jiangmaowei@sia.cn

Sciences

北京，中国

Beijing, China

科学

zengpengyu@sia.cn

Huan Liu

刘欢

Shenyang Institute of Automation,

沈阳自动化研究所，

Chinese Academy of Sciences

中国科学院

Shenyang, China

沈阳，中国

liuhuan@sia.cn

Wenbo Chen

陈文博

Shenyang Institute of Automation,

沈阳自动化研究所，

Chinese Academy of Sciences

中国科学院

Shenyang, China

沈阳，中国

University of Chinese Academy of University of Chinese Academy of

中国科学院大学

Sciences

科学学院

Beijing, China

北京，中国

chenwenbo@sia.cn Haoran Liu Shenyang Institute of Automation, Chinese Academy of Sciences Shenyang, China Sciences Beijing, China liuhaoran@sia.cn

chenwenbo@sia.cn 刘浩然 沈阳自动化研究所，中国科学院 沈阳，中国 科学学院 北京，中国 liuhaoran@sia.cn

# ABSTRACT

# 摘要

Time series forecasting is a long-standing challenge due to the real-world information is in various scenario (e.g., energy, weather, traffic, economics, earthquake warning). However some mainstream forecasting model forecasting result is derailed dramatically from ground truth. we believe it’s the reason that models’ lacking ability of capturing frequency information which richly contains in real world datasets. At present, the mainstream frequency information extraction methods are Fourier transform(FT) based. However, use of FT is problematic due to Gibbs phenomenon. If the values on both sides of sequences differ significantly, oscillatory approximations are observed around both sides and high frequency noise will be introduced. Therefore We propose a novel frequency enhanced channel attention that adaptively modelling frequency interdependencies between channels based on Discrete Cosine Transform which would intrinsically avoid high frequency noise caused by problematic periodity during Fourier Transform, which is defined as Gibbs Phenomenon. We show that this network generalize extremely effectively across six real-world datasets and achieve state-of-the-art performance, we further demonstrate that frequency enhanced channel attention mechanism module can be flexibly applied to different networks. This module can improve the prediction ability of existing mainstream networks, which reduces MSE on LSTM, on Reformer, on Informer, on Autoformer, on Transformer, etc., at a slight computational cost, with just a few line of code. Our codes and data are available at https://github.com/Zero-coder/FECAM.

时间序列预测是一个长期存在的挑战，因为现实世界的信息在不同的场景中（例如，能源、天气、交通、经济、地震预警）各不相同。然而，一些主流预测模型的预测结果与实际情况相去甚远。我们认为，这是因为模型缺乏捕捉频率信息的能力，而频率信息在现实世界数据集中丰富存在。目前，主流的频率信息提取方法基于傅里叶变换（FT）。然而，使用FT存在问题，因为吉布斯现象。如果序列两边的值相差很大，则在两边都会观察到振荡近似，并且会引入高频噪声。因此，我们提出了一种新颖的频率增强通道注意力机制，该机制基于离散余弦变换自适应地建模通道之间的频率依赖性，这可以本质上避免由于傅里叶变换中的问题周期性引起的高频噪声，这一定义为吉布斯现象。我们展示了这个网络在六个现实世界数据集上泛化效果极其有效，并取得了最先进的表现。我们进一步证明，频率增强通道注意力机制模块可以灵活地应用于不同的网络。这个模块可以提高现有主流网络的预测能力，在略微的计算成本下，通过仅仅几行代码，减少了LSTM上的 MSE，Reformer上的 ，Informer上的 ，Autoformer上的 ，Transformer上的 等。

# PVLDB Reference Format:

# PVLDB参考文献格式：

Maowei Jiang , Pengyu Zeng , Kai Wang\*, Huan Liu, Wenbo Chen, and Haoran Liu. FECAM: Frequency Enhanced Channel Attention Mechanism for Time Series Forecasting. PVLDB, 14(1): XXX-XXX, 2020. doi:XX.XX/XXX.XX

江茂伟 ，曾鹏宇 ，王凯\*，刘欢，陈文博，刘豪然。FECAM：用于时间序列预测的频率增强通道注意力机制。PVLDB，14(1)：XXX-XXX，2020。数字对象标识符：XX.XX/XXX.XX

# PVLDB Artifact Availability:

# PVLDB 艺术品可用性：

The source code, data, and/or other artifacts have been made available at https://github.com/Zero-coder/FECAM.

源代码、数据和其他相关艺术品已发布在 https://github.com/Zero-coder/FECAM。

# 1 INTRODUCTION

# 1 引言

Time series forecasting (TSF) enables decision-making with the estimated future evolution of metrics or events, thereby playing a crucial role in various scientific and engineering fields such as weather forecasting , estimation of future illness cases , energy consumption management , traffic flow , and financial investment , to name a few.

时间序列预测（TSF）能够通过预测指标或事件的未来演变来辅助决策，因此在诸如天气预报 ，未来病例估计 ，能源消耗管理 ，交通流量 和金融投资 等多个科学和工程领域发挥着至关重要的作用。

With the growing data availability and computing power in recent years, it is shown that deep learning-based TSF methods can achieve much better prediction performance than traditional approaches [21].

近年来，随着数据可用性和计算能力的增长，研究表明基于深度学习的时间序列预测方法比传统方法 [21] 的预测性能要优越得多。

In recent year, Transformers [35] have achieved progressive breakthrough on extensive areas . Especially in time series forecasting, credited to their stacked structure and the capability of attention mechanisms, Transformers [18, 35, 47] can naturally capture the temporal dependencies among time points, thereby fitting the series forecasting task perfectly.

近年来，Transformer [35] 在广泛的领域取得了逐步的突破 。特别是在时间序列预测方面，得益于其堆叠结构和对注意力机制的强大能力，Transformer [18, 35, 47] 能够自然地捕捉时间点之间的时间依赖性，从而完美地适应序列预测任务。

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit https://creativecommons.org/licenses/by-nc-nd/4.0/ to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

本作品遵循知识共享 BY-NC-ND 4.0 国际许可。访问 https://creativecommons.org/licenses/by-nc-nd/4.0/ 查看该许可的副本。对于本许可未涵盖的任何用途，请通过发送电子邮件至 info@vldb.org 获取许可。版权由所有者/作者持有。出版权授权给 VLDB Endowment。

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.

《VLDB基金会会议录》，第14卷，第1期 ISSN 2150-8097。

doi:XX.XX/XXX.XX

doi:XX.XX/XXX.XX

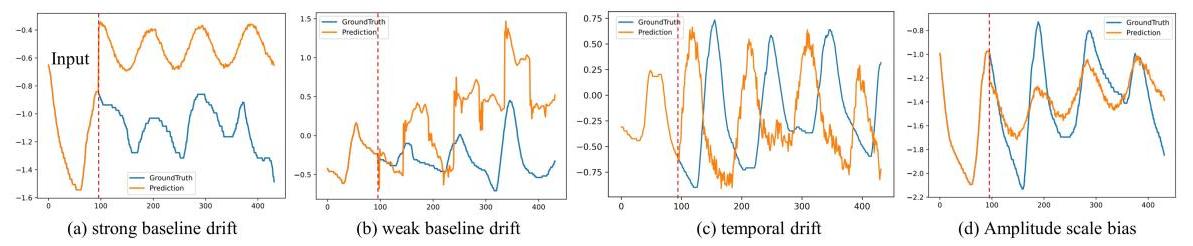


Figure 1: The discrepancy between ground truth and forecasting output on real-world dataset ETTm2,(a) is from the vanilla LSTM,(b)is from vanilla Informer(c)is from vanilla transformer (d)is from vanilla autoformer.

图1：真实世界数据集ETTm2上的地面真实值与预测输出之间的差异，（a）来自标准LSTM，（b）来自标准Informer，（c）来自标准transformer，（d）来自标准autoformer。

Despite the promising results of TSF methods, we found that the prediction of those methods, like transformers and LSTM is way derailed from the distribution of the ground truth of datasets, such as baseline drift in Fig.1(a) and temporal drift in Fig.1(c), we believe it’s the reason that models’ lacking ability of capturing frequency information which richly contains in real world datasets(Fig.2), Therefore, the thing is that there still have room for improvement for these TSF mainstream methods to exploiting the natural property of time series data what we call frequency during modeling.

尽管时间序列预测（TSF）方法取得了有希望的结果，但我们发现这些方法的预测，如transformers和LSTM，与数据集的地面真实值的分布相去甚远，例如图1（a）中的基线漂移和图1（c）中的时间漂移，我们认为这是模型缺乏捕获富含在真实世界数据集中的频率信息的能力的原因。因此，这些TSF主流方法在建模过程中利用时间序列数据的自然属性，即我们所说的频率，仍有改进的空间。

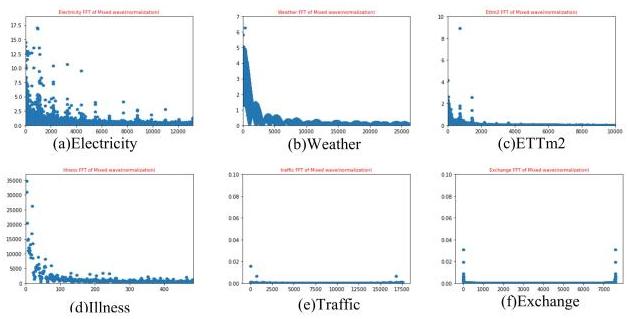


Figure 2: six real world datasets visualization in Frequency domain, we can see most energy is contained in low frequency range.

图2：六个真实世界数据集在频率域的可视化，我们可以看到大部分能量包含在低频范围内。

Some efforts has been done for getting frequency representation and reconstructing temporal signal based on Fourier Transform and it’s inverse transform. However, Fourier Transform (FT) would introduce high-frequency components for its problematic periodity, causing error value for boundary information which call Gibbs phenomenon and new round of computation consumption for inverse operation for avoiding complex operation in networks. Unlike FT/IFT based methods, Our method is based on Discrete Cosine Transform which would intrinsically eradicate Gibbs Phenomenon mentioned above and save unnecessary consumption of inverse transform, and for better exploiting utility of relationship between different time-series variate, we propose Frequency Enhanced Channel Attention Mechanism as a general framework, which empowers Transformer-based method and other mainstream models like LSTM, with better predictive ability for real-world time series.Consequently, effectively utilizing frequency information of time series enable us to perform forecasting with reasonable accuracy. Our method achieves state-of-the-art performance on six real-world benchmarks as a model. Furthermore, as a module FE-CAM can generalize to various Networks for further improvement, with just few line codes.

已经进行了一些努力来获取频率表示和基于傅里叶变换及其逆变换重构时间信号。然而，由于傅里叶变换（FT）存在周期性问题，它会引入高频成分，导致边界信息的误差值，这被称为吉布斯现象，并且为了避免网络中的复杂运算，逆操作需要新的计算消耗。与基于FT/IFT的方法不同，我们的方法基于离散余弦变换，这会内在地消除上述的吉布斯现象，并节省逆变换的不必要消耗。为了更好地利用不同时间序列变量之间的关系，我们提出了频率增强通道注意力机制作为一个通用框架，它赋予了基于Transformer的方法以及其他主流模型（如LSTM）更好的实际时间序列预测能力。因此，有效地利用时间序列的频率信息使我们能够以合理的精度进行预测。我们的方法在六个现实世界基准测试中作为模型实现了最先进的表现。此外，作为模块，FE-CAM可以推广到各种网络以进行进一步的改进，只需几行代码即可。

To this end, we propose a general feature extraction method for sequence modeling and forecasting, named frequency enhanced channel attention mechanism, which intrinsically eradicate Gibbs Phenomenon caused by Fourier Transform for the first time in time series forecasting.Our method achieves state-of-the-art performance on six real-world datasets, and can be generalized to other model architectures with just few line codes. The contributions of this paper are summarized as follows:

为此，我们提出了一种用于序列建模和预测的通用特征提取方法，名为频率增强通道注意力机制，它内在地消除了由傅里叶变换引起的时间序列预测中的吉布斯现象。我们的方法在六个现实世界数据集上实现了最先进的表现，并且可以用几行代码推广到其他模型架构。

* We theoretically prove that our method can mitigate Gibbs phenomenon which would introduce high frequency noise during Fourier Transform, and we demonstrate that GAP is the lowest frequency component of DCT.
* 我们从理论上证明了我们的方法可以减轻傅里叶变换过程中引入的高频噪声的吉布斯现象，并且我们证明了GAP是DCT的最低频率成分。
* Based on above proof, We build the channel attention in frequency domain and propose our method with frequency enhanced channel mechanism for time-series forecasting. For generalization, we generalize frequency-enhanced channel attention into module that can be easily and flexibly adapted into other mainstream time series forecasting models to get better performance on six real-world datasets.
* 基于上述证明，我们在频率域中构建了通道注意力，并提出了带有频率增强通道机制的时间序列预测方法。为了泛化，我们将频率增强通道注意力推广为一个模块，它可以轻松灵活地适应其他主流时间序列预测模型，以在六个现实世界数据集上获得更好的性能。
* Extensive experiments on various TSF datasets show that FECAM as a general method consistently boosts four mainstream Transformers and non-transformer based methods like LSTM by a considerable margin and achieves state-of-the-art performance on six real-world datasets.
* 在各种时间序列预测（TSF）数据集上的广泛实验表明，作为一种通用方法，FECAM一致地提高了四种主流的Transformer模型以及基于LSTM等非Transformer方法的性能，提高了相当大的幅度，并在六个现实世界数据集上取得了最先进的表现。

# 2 RELATED WORK AND PRELIMINARY

# 2 相关工作与预备知识

# 2.1 Deep Learning Models for Times series forecasting

# 2.1 深度学习模型在时间序列预测中的应用

In recent years, deep learning models with meticulously designed architectures have achieved excellent progress in TSF tasks. RNN-based models are proposed for application in an auto-regressive manner for sequence modeling, but the recurrent structure can suffer from problem of modeling long-term dependency. Shortly afterwards, Transformer [35] emerges and shows great power in sequence modeling and gains great achievements in various downstream tasks. To solve the quadratic computation consumption on sequence length, subsequent works aim to decrease Self-Attention’s complexity. Particularly in long-term time series forecasting, Informer [47] extends Self-Attention with KL-divergence criterion to select dominant queries. Reformer [18] introduces local-sensitive hashing (LSH) mechanism to approximate attention by allocated similar queries. Not just improvement of reduction complexity, the following models further develop delicate building blocks for time series forecasting. Autoformer [40] coalesce the decomposition blocks into a canonical structure and designs Auto-Correlation to capture series-wise connections. Pyraformer [22] designs pyramid attention module (PAM) to capture temporal dependencies with different hierarchies. Transformer-based models have taken the place of RNN-based models in almost all sequence modeling tasks, thanks to the effectiveness and efficiency of the self-attention mechanisms. Various Transformer-based TSF methods are proposed in the literature. These works typically focus on the challenging long-term time series forecasting problem, taking advantage of their remarkable long sequence modeling capabilities.

近年来，深度学习模型通过精心设计的架构在时间序列预测（TSF）任务上取得了卓越的进展。基于RNN的模型 被提出用于以自回归方式对序列进行建模，但循环结构在建模长期依赖性方面可能存在问题。不久之后，Transformer [35] 出现，并在序列建模方面显示出强大的能力，在各种下游任务中取得了巨大成就。为了解决序列长度上的二次计算消耗问题，后续工作旨在降低自注意力的复杂度。特别是在长期时间序列预测中，Informer [47] 通过引入KL散度标准选择主导查询来扩展自注意力。Reformer [18] 引入局部敏感哈希（LSH）机制，通过分配相似的查询来近似注意力。不仅仅是降低复杂度的改进，后续模型进一步为时间序列预测开发了精细的建筑块。Autoformer [40] 将分解块合并为规范结构，并设计Auto-Correlation来捕捉序列间的连接。Pyraformer [22] 设计了金字塔注意力模块（PAM），以捕捉不同层次的时间依赖性。由于自注意力机制的有效性和效率，基于Transformer的模型几乎在所有序列建模任务中取代了基于RNN的模型。文献中提出了各种基于Transformer的时间序列预测方法。这些工作通常关注于具有挑战性的长期时间序列预测问题，利用它们在长序列建模方面的显著能力。

Although the transformers can capture long-range dependency in the time domain, it does not explicitly model the pattern occurrences in the frequency domain that plays an important role in tracking and predicting data points over various time cycles.

虽然Transformer能够捕获时间域中的长距离依赖性，但它并没有显式地建模频率域中的模式发生，这在跟踪和预测不同时间周期上的数据点中起着重要的作用。

Different from previous works focusing on architectural design based on transformers, we analyze the series forecasting task from the natural view of frequency, which is the essential property of time series.It is also notable that as a general block, our proposed frequency-enhanced channel block can be easily applied to various models with a few operation. In the following subsection, we highlight our insights and motivate our work.

与之前专注于基于变压器的架构设计的工作不同，我们从频率的自然视角分析序列预测任务，频率是时间序列的本质属性。值得注意的是，作为通用模块，我们提出的频率增强通道模块可以轻松应用于各种模型，只需少量操作。在接下来的子节中，我们突出我们的见解并阐述我们的工作动机。

# 2.2 Frequency Representation for time series forecasting

# 2.2 时间序列预测的频率表示

Frequency is an indispensable information of time series, and real world datasets often contain rich frequency information as shown in Fig.2, which allows better utilization of the capabilities of deep learning models. To utilize frequency information, Auto-former [40] use FFT in efficient computing of auto-correlation function, FNO [14] is used as an inner block of networks to perform representation learning in low-frequency domain, DCTnet [42] use Discrete Cosine Transform to compress information for keeping more original picture information in CV task. Most of these work was based on Fourier Transform which is helpful for extracting frequency features. However, most of the FT-based methods use Fourier Transform to get the frequency information and use Inverse Fourier Transform to reconstruct temporal information for avoiding complex-number training, which introduces new amount of computation, which is avoidable if using DCT for time-frequency transformation, what’s more the implicit periodicity of DFT gives rise to boundary discontinuity that result in significant high-frequency content which is known as Gibbson Phenomeneon. After quantization, Gibbs Phenomeneon causes the boundary points to take on erraneous values. SENET [15] only use GAP which is the lowest component of DFT and DCT for channel representation, meaning discarding other frequency-component information.

频率是时间序列不可或缺的信息，如图2所示，现实世界的数据集通常包含丰富的频率信息，这允许更有效地利用深度学习模型的能力。为了利用频率信息，Auto-former [40] 在自相关函数的高效计算中使用了FFT，FNO [14] 被用作网络的内块，在低频域进行表示学习，DCTnet [42] 使用离散余弦变换来压缩信息，以在计算机视觉任务中保留更多原始图像信息。这些工作中的大多数是基于傅里叶变换的，这有助于提取频率特征。然而，大多数基于FT的方法使用傅里叶变换获取频率信息，并使用逆傅里叶变换重建时间信息以避免复数训练，这引入了新的计算量，如果使用DCT进行时频变换，则可以避免这种情况。更重要的是，DFT的隐含周期性导致了边界不连续，从而产生了显著的高频内容，这就是所谓的吉布斯现象。量化后，吉布斯现象导致边界点取错误的值。SENET [15] 只使用GAP，即DFT和DCT的最低分量进行通道表示，这意味着丢弃了其他频率分量信息。

# 2.3 Problem of Gibbs Phenomenon

# 2.3 吉布斯现象的问题

The Gibbs phenomenon involves both the fact that Fourier sums overshoot at a jump discontinuity, and that this overshoot does not die out as more sinusoidal terms are added. And this would cause high frequencies noise which is supposed to be avoidable for time series forecasting. We demonstrate the phenomenon for square wave (In Fig.3) with the additive synthesis of a square wave with an increasing number of harmonics. The Gibbs phenomenon is visible especially when the number of harmonics is large. We give the mathematic description of Gibbs Phenomenon below.

吉布斯现象 包括傅里叶和式在跳跃不连续点处的超射事实，以及随着更多正弦项的加入这种超射不会消失。这会导致时间序列预测中应避免的高频噪声。我们通过正方形波的谐波数目增加的叠加演示了这一现象（见图3）。当谐波数目较大时，吉布斯现象尤为明显。下面我们给出吉布斯现象的数学描述。

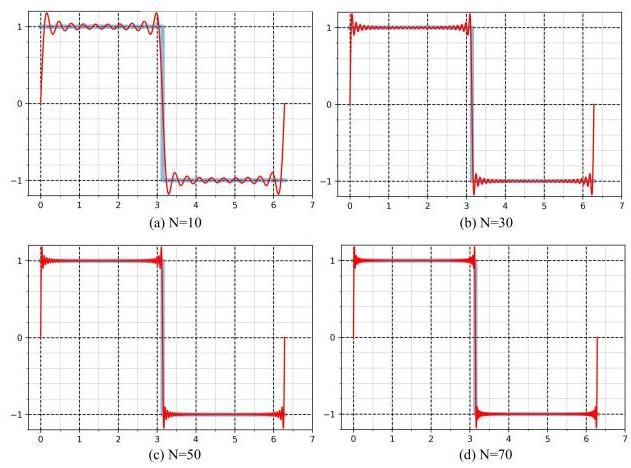


Figure 3: Gibbs Phenomenon with increasing harmonics component.

图3：随着谐波分量增加的吉布斯现象。

Formal mathematical description of the phenomenon: Let be a piecewise continuously differentiable function which is periodic with some period . Suppose that at some point , the left limit and right limit of the function differ by a non-zero jump of :

现象的正式数学描述：设 是一个分段连续可微的函数，且具有某个周期 的周期性。假设在某个点 ，函数 的左极限 和右极限 之间的非零跳跃为 ：

For each positive integer , let be the th partial Fourier series

对于每个正整数 ，设 为 阶的部分傅里叶级数

where the Fourier coefficients are given by the usual formulae

其中傅里叶系数 由通常的公式给出

Then we have:

然后我们有：

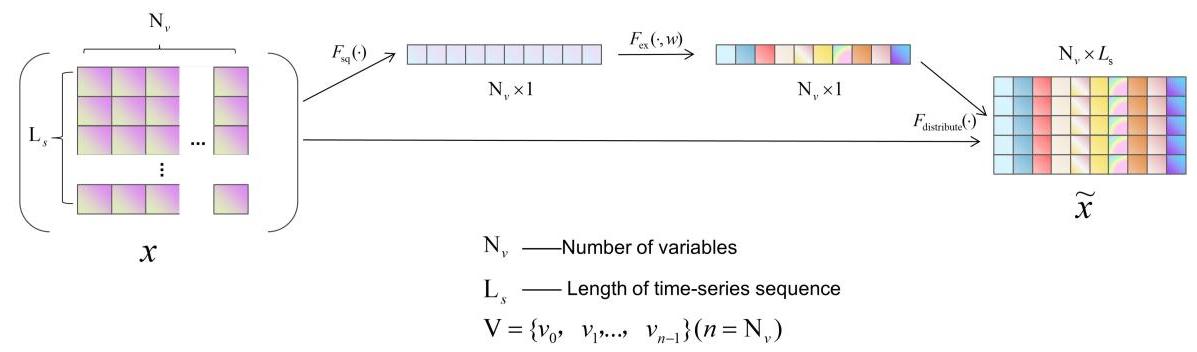


Figure 4: SENET channel attention(Squeeze and excitation Network), Fsq(.) represent usg1d-global average pooling to extract global information a full connected layer for from channel and redistribute weight for each channel.

图4：SENET通道注意力（挤压和激励网络），Fsq(.) 表示使用1d全局平均池化提取全局信息，一个全连接层用于来自通道的权重重新分配。

and

并且

but

但是

More generally, if is any sequence of real numbers which converges to as , and if the jump of is positive then

更一般地，如果 是任何收敛于 的实数序列，当 ，且如果 的跳跃是正的，那么

and

并且

If instead the jump of is negative, one needs to interchange limit superior with limit inferior, and also interchange the and signs, in the above two inequalities.

如果 的跳跃是负的，则需要在上面的两个不等式中交换上极限和下极限，并交换 和 的符号。

# 3 FECAM: FREQUENCY ENHANCED CHANNEL ATTENTION MECHANISM

# 3 FECAM：频率增强通道注意力机制

Frequency is a natural auxiliary means to analyze time series. It is important and intuitive to introduce frequency information into time series models. However, most time series models tend to ignore the impact of frequency information on time series tasks, resulting in failure to learn the inherent characteristics of time series information. Most methods of extracting frequency information are based on FT and IFT, However, methods based on FT and IFT tends to introduce high-frequency noise due to problematic peri-odity which is known as Gibbs Phenomenon, Frequency Enhanced Frequency Channel Attention Mechanism can intrinsically avoid problem mentioned above and automatically acquire the importance of each channel through learning, it also suppresses features that are not useful for the current task.

频率是分析时间序列的一种自然辅助手段。将频率信息引入时间序列模型是重要且直观的。然而，大多数时间序列模型倾向于忽略频率信息对时间序列任务的影响，导致无法学习时间序列信息的内在特性。大多数提取频率信息的方法基于傅里叶变换（FT）和逆傅里叶变换（IFT），但是基于FT和IFT的方法由于周期性问题引入了高频噪声，这被称为吉布斯现象。频率增强的频率通道注意力机制可以从本质上避免上述问题，并通过学习自动获取每个通道的重要性，它还抑制了对当前任务无用的特征。

We expect the learning of channel interdependencies features to be enhanced by explicitly modelling in frequency domain.

我们期望通过在频率域中显式建模来增强通道相互依赖特征的学习。

# 3.1 Channel Attention and DCT

# 3.1 通道注意力和DCT

We first elaborate on the definitions of discrete cosine transform and channel attention mechanism.

我们首先详细阐述离散余弦变换和通道注意力机制的定义。

3.1.1 Revisiting Channel attention. The channel attention mechanism has been successfully introduced to CNNs. Squeeze-and-excitation (SE) block [15] models the interdependencies between the channels of feature maps with global information and recalibrate the feature maps to improve representation ability. It consists of squeeze and excitation two steps which are depicted in Fig .4. For time-series signals is the number of channels, is the length of the temporal sequence, this type of tensor could be anywhere in the time-series model.

3.1.1 重新审视通道注意力。通道注意力机制已经被成功引入到卷积神经网络（CNNs）中。挤压和激发（SE）块[15]使用全局信息来建模特征图通道之间的相互依赖，并重新校准特征图以提高表示能力。它包括挤压和激发两个步骤，如图4所示。对于时间序列信号 是通道数， 是时间序列的长度，这种类型的张量可以位于时间序列模型中的任何位置。

For temporal signals, the squeeze step applies GAP on temporal dimension to generate channel wise descriptor. Officially, a statistic is generated by shrinking through its temporal dimension , such that the -th item of is calculated by:

对于时间信号，挤压步骤在时间维度上应用全局平均池化（GAP）以生成通道描述符。正式来说，一个统计量 通过收缩 在其时间维度 上生成，使得 的第 个元素是通过以下方式计算的：

Where represent the channel, and temporal dimension respectively. The scalar is the -th element of , Then the excitation step aims to modelling channel-wise dependencies by using two fully-connected layers and with a bottleneck architecture and non-linearity:

其中 分别代表通道和时序维度。标量 是 的第 个元素，然后激励步骤旨在通过使用具有瓶颈架构和非线性的两个全连接层 和 来建模通道间的依赖关系：

where at is the learned attention vector which dot multiplies to the original feature map to re-scale each channel, and refer to ReLU and sigmoid activation function respectively.

其中在 位置的是学习到的注意力向量，它与原始特征图逐点相乘以重新缩放每个通道， 和 分别指代 ReLU 和 sigmoid 激活函数。

3.1.2 Frequency representation for time series. Sometimes, frequency information contains more information that can be found, but it is difficult to mine in the time domain. For example, when a signal is disturbed by noise, its waveform will become messy, Or we can’t distinguish it from noise in time domain. But it can be clearly distinguished from the frequency domain. Instead of well-known Fourier Transform, Our method introduce frequency information by Discrete Cosine Transform which can intrinsically avoid G-phenomenon and inverse transform operation.

3.1.2 时间序列的频率表示。有时，频率信息包含的信息比时域中可找到的更多，但在时域中很难挖掘。例如，当一个信号被噪声干扰时，其波形会变得杂乱无章，或者在时域中无法将其与噪声区分开来。但是，在频率域中可以清晰地将其区分开来。我们的方法通过离散余弦变换（Discrete Cosine Transform）引入频率信息，这可以内在地避免 G-现象和逆变换操作。

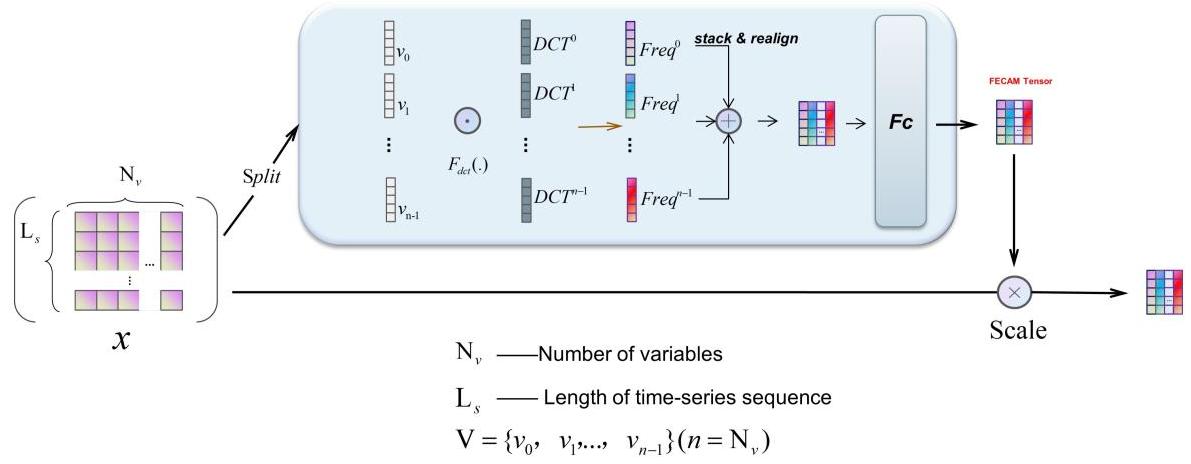


Figure 5: Structure of Frequency Enhanced Channel Attention Mechanism. split every sequence of multivariate time series in each channel and Fdct(.) stands for Discrete Cosine Transform, stack&realign each channel together.

图 5：增强频率通道注意力机制的结构。在每个通道中拆分多变量时间序列的每个序列，Fdct(.) 表示离散余弦变换，将每个通道堆叠并重新对齐。

Discrete Cosine Transform (DCT)

离散余弦变换（DCT）

Typically, the basis function of one-dimensional (1D) DCT is:

通常，一维（1D）DCT 的基函数是：

Then the 1D DCT can be written as:

然后，一维 DCT 可以写成：

s.t. , In which is the 1D DCT frequency spectrum, is the input, is the length of . Correspondingly, the inverse 1D DCT can be written as:

使得 ，在其中 是一维 DCT 频率谱， 是输入， 是 的长度。相应地，逆一维 DCT 可以写成：

s.t. , In which , Please note that in Eqs. 2 and 3, some constant normalization factors are removed for simplicity, which will not affect the results in this work.

令 ，其中 ，请注意，在公式2和3中，为了简化，省略了一些常数归一化因子，这不会影响本文的结果。

# 3.2 Frequency Enhanced Channel Attention Mechanism

# 3.2 频率增强通道注意力机制

In this section, we first theoretically discuss the problem of existing channel attention mechanisms. Based on the theoretical analysis, we then elaborate on the network design of the proposed method.

在本节中，我们首先从理论上讨论现有通道注意力机制的问题。基于理论分析，然后我们详细阐述所提出方法的网络设计。

Although GAP is a widely used operation in many attention mechanism as a standard squeezing method, we argue that simply use average-pooling on temporal dimension cause inadequate information extraction from time series which would even leads to information loss. Since GAP is the lowest frequency component of DCT and DFT, we mitigate this problem by introducing more frequency information. Rather than DFT, We use DCT to evade the Gibbs phenomenon mentioned many times before.

尽管 GAP 是许多注意力机制中作为标准压缩方法广泛使用的操作，但我们认为仅在时间维度上使用平均池化会导致从时间序列中提取的信息不足，甚至导致信息丢失。由于 GAP 是 DCT 和 DFT 的最低频率分量，我们通过引入更多频率信息来缓解这个问题。我们使用 DCT 而不是 DFT，以避免之前多次提到的吉布斯现象。

Theorem 1. 1d-GAP is a lowest component of 1D DCT, and its result is proportional to the lowest frequency component of -DCT.

定理1。1d-GAP 是 1D DCT 的最低分量，其结果与 -DCT 的最低频率分量成比例。

In Eq.14, represents the lowest frequency component of 1D DCT, and it is proportional to GAP. In this way, Theorem 1 is proved.

在公式14中， 表示 1D DCT 的最低频率分量，它与 GAP 成比例。这样，定理1得到了证明。

According to Theorem 1, without any surprise, we can sure that using GAP for feature extraction in channel attention means only the lowest frequency in obtained. All other frequency components are ignored, which supposed to be included in presenting channels.

根据定理1，毫不意外，我们可以确信在通道注意力中使用 GAP 进行特征提取意味着仅获得了最低频率。所有其他频率分量都被忽略，而这些分量本应包含在当前通道中。

Theorem 2. Discrete Cosine Transform can intrinsically avoid Gibbs Phenomenon caused by periodic problem of Discrete Fourier Transform and Inverse Discrete Fourier Transform, and have a more efficient energy compaction than Fourier Transform.

定理2。离散余弦变换（DCT）本质上可以避免由离散傅里叶变换（DFT）和逆离散傅里叶变换引起的吉布斯现象，并且比傅里叶变换具有更有效的能量压缩。

Discrete Cosine Transform is actually the DFT whose input signal is a real even function (proved in Derivation of DCT in Appendix A). Since Discrete Cosine Transform is using symmetric expansion for it’s periodic extension (In Fig.6). Therefore, followed with eq.1, we have:

离散余弦变换实际上是输入信号为实偶函数的DFT（附录A中已证明DCT的推导）。由于离散余弦变换对其周期扩展使用对称展开（如图6所示），因此，根据等式1，我们有：

Equation 15 means that there’s no jump discontinuity which is necessary condition of Gibbs Phenomenon.

公式15意味着不存在跳变不连续性，这是吉布斯现象的必要条件。

Then follow with eq. 4 and eq.5, then we have:

然后根据等式4和等式5，我们有：

Follow with eq. 7 and eq. 8,

根据等式7和等式8，

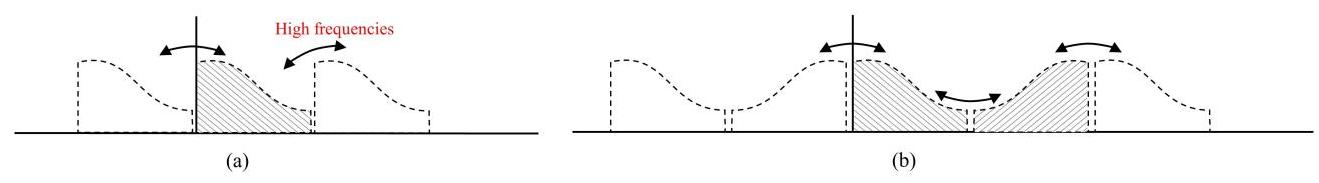


Figure 6: Extension of Discrete Fourier Transform and Discrete Cosine Transform.

图6：离散傅里叶变换和离散余弦变换的扩展。

As we can see, the limit of the formula converges at this point, no oscillation observed, thus fundamentally eliminating the Gibbs effect. Because IFT and FT is consistent in mathematical nature, it is also true for the inverse DFT. It’s worthy to note that in the DFT case the periodic extension introduces discontinuities, which not happen for the DCT due to its property of symmetry extension, it is the elimination of this artificial discontinuity which contains a lot of high frequencies make the DCT is much more energy efficient than Discrete Fourier Transform. To this extent, Theorem 2 is proved. And we have done experiments to validate our Theorem 2 in section 4.4.

由此可见，公式的极限在此点收敛，没有观察到振荡，从而根本消除了吉布斯效应。由于逆傅里叶变换和傅里叶变换在数学本质上是一致的，对于逆DFT也适用。值得注意的是，在DFT的情况下，周期扩展引入了不连续性，这不会在DCT中出现，因为DCT具有对称扩展的性质，正是消除了这种包含大量高频的人工不连续性，使得DCT比离散傅里叶变换具有更高的能量效率。至此，定理2得证。我们已在第4.4节进行了实验验证我们的定理2。

According to Theorem 2, we can found that in the DFT case the extension introduces discontinuities and this does not happen for DCT, due to the symmetry of its periodic extension, then our method eliminate this artificial discontinuity which contains a lot of high frequencies.

根据定理2，我们可以发现，在DFT的情况下，扩展引入了不连续性，而这不会在DCT中出现，因为其周期扩展具有对称性，然后我们的方法消除了这种包含大量高频的人工不连续性。

For capturing more time series information from feature map, we try to introduce DCT for getting more frequency components instead of only the GAP for lowest frequency [15]. Since DCT weight are constant, it can be pre-calculated only once and saved in advance, what’s more, results are real number, which means no training time for inverse transform and number of network parameters. Therefore, we propose frequency enhanced channel attention mechanism (FECAM) which can not only be used as a model for forecasting with just adding a projection layer but also can be seamlessly added to the existing time series forecasting models for improving their prediction performance. The overall structure of FECAM is shown in Fig.5.

为了从特征图捕获更多的时间序列信息，我们尝试引入DCT来获取更多频率分量，而不仅仅是最低频率的GAP [15]。由于DCT权重是常数，它只需要预先计算一次并保存起来，更重要的是，结果为实数，这意味着逆变换不需要训练时间，并且网络参数的数量也不会增加。因此，我们提出了频率增强通道注意力机制（FECAM），它不仅可以作为仅通过添加一个投影层即可用于预测的模型，还可以无缝地添加到现有的时间序列预测模型中，以提高它们的预测性能。FECAM的整体结构如图5所示。

First, FECAM splits the input feature maps along the channel dimension into sub-groups as , in which , Subsequently, for sub-group will be processed by a corresponding DCT frequency component ranging from low frequency to high frequency, Every single channel will processed by the same frequency component, In this way we have:

首先，FECAM沿着通道维度将输入特征图分为 个子组，即 ，其中 ，随后，每个子组将由相应的DCT频率分量处理，范围从低频到高频，每个单独的通道都将处理相同的频率分量，这样我们得到：

s.t. , in which are the frequency component 1D indices corresponding to , and is the dimensional vector after the discrete cosine transformation. The whole frequency channel vector can be obtained by stack operation.

其中 ， 是对应于 的频率分量1D索引， 是离散余弦变换后的 维向量。整个频率通道向量可以通过堆叠操作获得。

In which Freq is the attention vector for . Once we obtain Freq, the attention weight can be learned through neural structure as SE-block. The whole frequency enhanced channel attention mechanism framework can be written as:

其中 Freq 是 的注意力向量。一旦我们获得Freq，注意力权重可以通过类似于SE块的神结构学习得到。整个频率增强通道注意力机制框架可以写成：

By doing so, each channel features interact with every frequency components to acquire important temporal information comprehensively from frequency domain, which would encourages networks to enhance the diversity of extracted features. In the subsequent experiment section 4.3 , we visualize the frequency channel attention tensor Fig.10, demonstrating that FECAM learned the importance of different channels in the frequency domain and the importance of different frequency component pairs in each channel.

通过这种方式，每个通道特征与每个频率分量相互作用，从频率域全面获取重要的时间信息，这将鼓励网络增强提取特征的多样性。在随后的实验部分4.3中，我们可视化了频率通道注意力张量图10，证明了FECAM在频率域中不同通道的重要性以及每个通道中不同频率分量对的重要性。

# 4 EXPERIMENTS

# 4 实验部分

We conduct extensive experiments to evaluate the performance of frequency enhanced channel mechanism network on six real-world time series forecasting benchmarks and further validate the generality of the proposed method on various mainstream Transformer variants and non-transformer based models.As a module embedding to other Networks, we have also done experiment of parameters increment and performance promotion and the visualization of frequency channel attention tensor to prove proposed method’s effectiveness and efficiency.

我们进行了广泛的实验，以评估频率增强通道机制网络在六个现实世界时间序列预测基准上的性能，并进一步验证了所提出方法在各种主流Transformer变体和非Transformer基于模型的通用性。作为一个嵌入到其他网络的模块，我们还进行了参数增加和性能提升的实验，以及频率通道注意力张量的可视化，以证明所提出方法的有效性和效率。

Datasets: Here are the descriptions of the datasets:

数据集：以下是数据集的描述：

Electricity : records the hourly electricity consumption of 321 clients from 2012 to 2014.

电能 ：记录了2012年至2014年321个客户的每小时电力消耗。

: contains the time series of oil temperature and power load collected by electricity transformers from July 2016 to July 2018. ETTm1/ETTm2 are recorded every 15 minutes, and ETTh1/ETTh2 are recorded every hour.

：包含从2016年7月至2018年7月电力变压器收集的油温和电力负荷的时间序列。ETTm1/ETTm2每15分钟记录一次，ETTh1/ETTh2每小时记录一次。

Exchange : collects the panel data of daily exchange rates from 8 countries from 1990 to 2016.

交易所 ：收集了从1990年至2016年8个国家的每日汇率的面板数据。

The Electricity dataset was acquired at https://archive.ics.uci.edu/ml/datasets/Electric-ityLoadDiagrams20112014

电能数据集可在 https://archive.ics.uci.edu/ml/datasets/Electric-ityLoadDiagrams20112014 获取

The ETT dataset was acquired at https://github.com/zhouhaoyi/ETDataset

ETT数据集可在 https://github.com/zhouhaoyi/ETDataset 获取

The Exchange dataset was acquired at https://github.com/thuml/Autoformer

交易所数据集可在 https://github.com/thuml/Autoformer 获取

Table 1: Statistics of datasets.

表1：数据集统计。

| Dataset | Variable Number | Sampling Frequency | Total Observations |
| --- | --- | --- | --- |
| Exchange | 8 | 1 Day | 7,588 |
| ILI | 7 | 1 Week | 966 |
| ETTm2 | 7 | 15 Minutes | 69,680 |
| Electricity | 321 | 1 Hour | 26,304 |
| Traffic | 862 | 1 Hour | 17,544 |
| Weather | 21 | 10 Minutes | 52,695 |

Table 2: Forecasting results comparison under different prediction lengths The input sequence length is set to 36 for ILI and 96 for the others.

表2：不同预测长度下的预测结果比较 输入序列长度设置为ILI的36，其他为96。

| Models | | Ours | | Autoformer | | Pyraformer | | Informer | | LogTrans | | Reformer | | LSTNet | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metric | | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| Exchange | 96 | 0.085 | 0.208 | 0.197 | 0.323 | 0.852 | 0.780 | 0.847 | 0.752 | 0.968 | 0.812 | 1.065 | 0.829 | 1.551 | 1.058 |
| 192 | 0.210 | 0.338 | 0.300 | 0.369 | 0.993 | 0.858 | 1.204 | 0.895 | 1.04 | 0.851 | 1.610 | 1.020 | 1.477 | 1.028 |
| 336 | 0.344 | 0.445 | 0.509 | 0.524 | 1.240 | 0.958 | 1.672 | 1.036 | 1.659 | 1.081 | 2.226 | 1.192 | 1.507 | 1.031 |
| 720 | 0.921 | 0.717 | 1.447 | 0.941 | 1.711 | 1.093 | 2.478 | 1.31 | 1.941 | 1.127 | 1.802 | 1.131 | 2.285 | 1.243 |
| ILI | 24 | 2.101 | 0.939 | 3.483 | 1.287 | 5.800 | 1.693 | 5.764 | 1.677 | 4.475 | 1.444 | 4.400 | 1.382 | 6.026 | 1.770 |
| 36 | 2.330 | 0.951 | 3.103 | 1.148 | 6.043 | 1.733 | 4.755 | 1.467 | 4.799 | 1.467 | 4.783 | 1.448 | 5.340 | 1.668 |
| 48 | 2.557 | 1.061 | 2.669 | 1.085 | 6.213 | 1.763 | 4.763 | 1.469 | 4.800 | 1.468 | 4.832 | 1.465 | 6.080 | 1.787 |
| 60 | 2.531 | 1.093 | 2.770 | 1.125 | 6.531 | 1.814 | 5.264 | 1.564 | 5.278 | 1.560 | 4.882 | 1.483 | 5.548 | 1.720 |
| ETTm2 | 96 | 0.188 | 0.275 | 0.255 | 0.339 | 0.409 | 0.479 | 0.365 | 0.453 | 0.768 | 0.642 | 0.658 | 0.619 | 3.142 | 1.365 |
| 192 | 0.265 | 0.336 | 0.281 | 0.340 | 0.673 | 0.641 | 0.533 | 0.563 | 0.989 | 0.757 | 1.078 | 0.827 | 3.154 | 1.369 |
| 336 | 0.318 | 0.362 | 0.339 | 0.372 | 1.210 | 0.846 | 1.363 | 0.887 | 1.334 | 0.872 | 1.549 | 0.972 | 3.160 | 1.369 |
| 720 | 0.416 | 0.417 | 0.422 | 0.419 | 4.044 | 1.526 | 3.379 | 1.388 | 3.048 | 1.328 | 2.631 | 1.242 | 3.171 | 1.368 |
| Electricity | 96 | 0.178 | 0.267 | 0.201 | 0.317 | 0.498 | 0.299 | 0.274 | 0.368 | 0.258 | 0.357 | 0.312 | 0.402 | 0.680 | 0.645 |
| 192 | 0.185 | 0.273 | 0.222 | 0.334 | 0.828 | 0.312 | 0.296 | 0.386 | 0.266 | 0.368 | 0.348 | 0.433 | 0.725 | 0.676 |
| 336 | 0.199 | 0.290 | 0.231 | 0.338 | 1.476 | 0.326 | 0.300 | 0.394 | 0.280 | 0.380 | 0.350 | 0.433 | 0.828 | 0.727 |
| 720 | 0.235 | 0.323 | 0.254 | 0.361 | 4.090 | 0.372 | 0.373 | 0.439 | 0.283 | 0.376 | 0.340 | 0.420 | 0.957 | 0.811 |
| Traffic | 96 | 0.493 | 0.318 | 0.613 | 0.388 | 0.684 | 0.393 | 0.719 | 0.391 | 0.684 | 0.384 | 0.732 | 0.423 | 1.107 | 0.685 |
| 192 | 0.496 | 0.319 | 0.616 | 0.382 | 0.692 | 0.394 | 0.696 | 0.379 | 0.685 | 0.390 | 0.733 | 0.420 | 1.157 | 0.706 |
| 336 | 0.511 | 0.325 | 0.622 | 0.337 | 0.699 | 0.396 | 0.777 | 0.420 | 0.733 | 0.408 | 0.742 | 0.420 | 1.216 | 0.730 |
| 720 | 0.547 | 0.343 | 0.660 | 0.408 | 0.712 | 0.404 | 0.864 | 0.472 | 0.717 | 0.396 | 0.755 | 0.423 | 1.481 | 0.805 |
| Weather | 96 | 0.182 | 0.242 | 0.266 | 0.336 | 0.354 | 0.392 | 0.300 | 0.384 | 0.458 | 0.490 | 0.689 | 0.596 | 0.594 | 0.587 |
| 192 | 0.223 | 0.281 | 0.307 | 0.367 | 0.673 | 0.597 | 0.598 | 0.544 | 0.658 | 0.589 | 0.752 | 0.638 | 0.560 | 0.565 |
| 336 | 0.270 | 0.320 | 0.359 | 0.395 | 0.634 | 0.592 | 0.578 | 0.523 | 0.797 | 0.652 | 0.639 | 0.596 | 0.597 | 0.587 |
| 720 | 0.338 | 0.374 | 0.419 | 0.428 | 0.942 | 0.723 | 1.059 | 0.741 | 0.869 | 0.675 | 1.130 | 0.792 | 0.618 | 0.599 |

ILI : collects the ratio of influenza-like illness patients versus the total patients in one week, which is reported weekly by Centers for Disease Control and Prevention of the United States from2002 and 2021.

ILI ：收集一周内流感样病例患者数与总患者数的比例，该数据由美国疾病控制与预防中心从2002年至2021年每周报告。

The ILI dataset was acquired at https://gis.cdc.gov/grasp/fluview/fluportaldashboar-d.html

ILI数据集来源于 https://gis.cdc.gov/grasp/fluview/fluportaldashboar-d.html

Traffic : contains hourly road occupancy rates measured by 862 sensors on San Francisco Bay area freeways from January 2015 to December 2016.

交通 ：包含2015年1月至2016年12月旧金山湾区高速公路上862个传感器测量的每小时道路占用率。

Weather : includes meteorological time series with 21 weather indicators collected every 10 minutes from the Weather Station of the Max Planck Biogeochemistry Institute in 2020.

天气 ：包括2020年马克斯·普朗克生物地球化学研究所气象站每10分钟收集的21个气象指标的时间序列。

The Traffic dataset was acquired at http://pems.dot.ca.gov/

交通数据集来源于 http://pems.dot.ca.gov/

The Weather dataset was acquired at https://www.bgc-jena.mpg.de/wetter/

天气数据集来源于 https://www.bgc-jena.mpg.de/wetter/

Table 3: Univariate results with different prediction lengths on datasets ETTm2 and Exchange. The input sequence length is set to 96 .

表3：不同预测长度 在数据集ETTm2和Exchange上的单变量结果。输入序列长度设置为96。

| Models | | Ours | | N-HiTs | | N-BEATS | | Autoformer | | Pyraformer | | Informer | | Reformer | | ARIMA | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metric | | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| Exchange | 96 | 0.085 | 0.208 | 0.114 | 0.248 | 0.156 | 0.299 | 0.241 | 0.387 | 0.290 | 0.439 | 0.591 | 0.615 | 1.327 | 0.944 | 0.112 | 0.245 |
| 192 | 0.206 | 0.338 | 0.250 | 0.387 | 0.669 | 0.665 | 0.273 | 0.403 | 0.594 | 0.644 | 1.183 | 0.912 | 1.258 | 0.924 | 0.304 | 0.404 |
| 336 | 0.344 | 0.445 | 0.434 | 0.516 | 0.611 | 0.605 | 0.508 | 0.539 | 0.962 | 0.824 | 1.367 | 0.984 | 2.179 | 1.296 | 0.736 | 0.598 |
| 720 | 0.759 | 0.672 | 1.061 | 0.773 | 1.111 | 0.860 | 0.991 | 0.768 | 1.285 | 0.958 | 1.872 | 1.072 | 1.280 | 0.953 | 1.871 | 0.935 |
| ETTm2 | 96 | 0.066 | 0.188 | 0.092 | 0.232 | 0.082 | 0.219 | 0.065 | 0.189 | 0.074 | 0.208 | 0.088 | 0.225 | 0.131 | 0.288 | 0.211 | 0.362 |
| 192 | 0.109 | 0.245 | 0.128 | 0.276 | 0,120 | 0.268 | 0.118 | 0.256 | 0.116 | 0.252 | 0.132 | 0.283 | 0.186 | 0.354 | 0.261 | 0.406 |
| 336 | 0.144 | 0.287 | 0.165 | 0.314 | 0.226 | 0.370 | 0.154 | 0.305 | 0.143 | 0.295 | 0.180 | 0.336 | 0.220 | 0.381 | 0.317 | 0.447 |
| 720 | 0.177 | 0.326 | 0.243 | 0.397 | 0.188 | 0.338 | 0.182 | 0.335 | 0.197 | 0.338 | 0.300 | 0.435 | 0.267 | 0.430 | 0.366 | 0.487 |

Table 1 summarizes overall statistics of the datasets. We follow the standard protocol that divides each dataset into the training, validation, and testing subsets according to the chronological order. The split ratio is for the ETT dataset and for others.

表1总结了数据集的整体统计数据。我们遵循标准协议，按照时间顺序将每个数据集分为训练集、验证集和测试集。分割比例为 对于ETT数据集，其他为 。

Baselines: We evaluate the single full-connected layer equipped by the Frequency Enhanced Channel Attention mechanism in both multivariate and uni-variate settings to demonstrate its effectiveness. For multivariate forecasting, we include six state-of-the-art deep forecasting models: Autoformer [40], Pyraformer [22], Informer [47], LogTrans [20], Reformer [18] and LSTNet [19]. For univariateforecasting, we include seven competitive baselines:N-HiTS [4], N-BEATS [27], Autoformer [40], Pyraformer [22], Informer [47], Reformer [18] and ARIMA [1]. In addition, we adopt the proposed framework on both the canonical and efficient variants of Transformers and classical RNNs:Transformer [35], Informer [47], Reformer [18] and Autoformer [40] and LSTM [12] to validate the generality of our framework.

基线：我们在多变量和单变量设置中评估了配备频率增强通道注意力机制的单一全连接层的效果，以展示其有效性。对于多变量预测，我们包括了六个最先进的深度预测模型：Autoformer [40]、Pyraformer [22]、Informer [47]、LogTrans [20]、Reformer [18] 和 LSTNet [19]。对于单变量预测，我们包括了七个有竞争力的基线：N-HiTS [4]、N-BEATS [27]、Autoformer [40]、Pyraformer [22]、Informer [47]、Reformer [18] 和 ARIMA [1]。此外，我们在Transformer和经典RNN的标准和高效变体上采用所提出的框架：Transformer [35]、Informer [47]、Reformer [18]、Autoformer [40] 和 LSTM [12]，以验证我们框架的通用性。

Implementation details: All the experiments are implemented in PyTorch [28] and conducted for three runs on a single NVIDIA GeForce RTX 309024GB GPU. Each model is trained by ADAM [17] using L2 loss with the initial learning rate of 10e-4 and batch size of 32. Each Transformer-based model contains two encoder layers and one decoder layer. We report the test MSE/MAE under different prediction lengths as the performance metric. A lower MSE/MAE indicates better performance of time series forecasting.

实施细节：所有实验都在PyTorch [28] 中实现，并在单个NVIDIA GeForce RTX 3090 24GB GPU上进行了三次运行。每个模型都通过ADAM [17] 使用L2损失进行训练，初始学习率为10e-4，批量大小为32。每个基于Transformer的模型包含两个编码器层和一个解码器层。我们报告了不同预测长度下的测试MSE/MAE作为性能指标。MSE/MAE值越低，表示时间序列预测性能越好。

# 4.1 Main Results

# 4.1 主要结果

Forecasting results: As for multivariate forecasting results, Our proposed method with a projection layer for forecasting achieves state-of-the-art performance in all benchmarks and prediction lengths (Table 2). Notably, Frequency Enhanced Channel Attention Mechanism outperforms other deep models impressively characterized by much less model parameters. Compared with Autoformer, the proposed FECAM yields an overall relative MSE reduction and relative MAE reduction. With the prediction length of 24 and 48, FECAM achieve an MSE reduction and respectively on ILI, with the the prediction length of 96,192,336,720, FECAM achieve relative MSE on Exchange compared to previous state-of-the-art results, which indicates that the potential of deep model is still constrained on ability of modelling in frequency domain. We also list the univariate results of two typical datasets with different frequency distribution (as shown in Fig.2). FECAM still realizes remarkable forecasting performance.

预测结果：关于多变量预测结果，我们提出的带有预测投影层的方法在所有基准测试和预测长度上均实现了最先进的表现（见表2）。值得注意的是，频率增强通道注意力机制在参数数量大幅减少的情况下，比其他深度模型表现更出色。与Autoformer相比，提出的FECAM实现了总体 相对均方误差减少和相对 平均绝对误差减少。在预测长度为24和48时，FECAM在ILI上分别实现了 均方误差减少 和 。在预测长度为96、192、336、720时，FECAM在交易所相比之前最先进的结果实现了 相对均方误差减少，这表明深度模型在频率域建模的能力上仍然存在潜在限制。我们还列出了两个具有不同频率分布的典型数据集的单变量结果（如图2所示）。FECAM仍然实现了显著的预测性能。

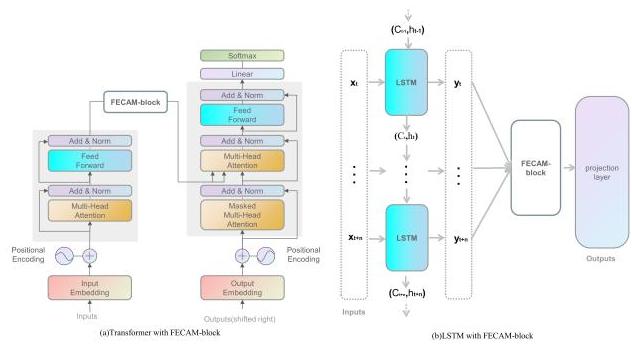


Figure 7: FECAM as a module embedded into other Networks, figure (a) represent module is put behind the encoder of Transformers and figure (b) represent module is put between LSTM output layer and projection layer.

图7：FECAM作为嵌入到其他网络中的模块，图(a)表示模块被放置在Transformers编码器之后，图(b)表示模块被放置在LSTM输出层和投影层之间。

Module generality: We apply our proposed method to four mainstream Transformers (as shown in Fig.7(a)) and a mainstream recurrent neural network LSTM (as shown in Fig.7(b)) and report the performance promotion of each model (Table 5). Our method consistently improves the forecasting ability of different models. Overall, it achieves averaged promotion on LSTM, on Reformer, on Informer, on Autoformer and on Transformer, making each of them surpass previous state-of-the-art. Compared to vanilla models, only a few parameters are increased by applying our method (See Table 4), and thereby their computational complexities can be preserved. It validates that Frequency Enhanced Channel Attention Mechanism is an effective and lightweight tool that can be widely applied to Transformer-based models and RNNs with a few line code, and enhances their ability of modelling in frequency domain to achieve state-of-the-art performance.

模块通用性：我们将我们提出的方法应用于四种主流的变换器（如图7(a)所示）以及一种主流的循环神经网络LSTM（如图7(b)所示），并报告了每个模型性能的提升（表5）。我们的方法一致地提高了不同模型的预测能力。总体而言，它在LSTM上实现了平均 的提升，在Reformer上为 ，在Informer上为 ，在Autoformer上为 ，在Transformer上为 ，使它们每一个都超过了之前的最先进水平。与原始模型相比，应用我们的方法仅增加了少数参数（见表4），因此它们的计算复杂度得以保持。这验证了频率增强通道注意力机制是一种有效且轻量级的工具，可以广泛地应用于基于变换器的模型和仅通过几行代码增强的RNN，并提高它们在频率域中的建模能力，以达到最先进的表现。

Table 4: Parameters increment and performance promotion of FECAM

表4：FECAM的参数增加和性能提升

| Models | LSTM | Reformer | Informer | Autoformer | Transformer |
| --- | --- | --- | --- | --- | --- |
| Vanilla |  |  | 11.33MB | 10.54MB | 10.54MB |
| Vanilla+Ours |  |  | 11.39MB | 10.69MB | 10.60MB |
| Parameters increment |  |  |  |  |  |
| Performance promotion | 35.99% | 10.01% | 8.71% | 8.29% |  |

Table 5: Performance promotion by applying our proposed method method to Transformers and RNNs.We report the averaged MSE/MAE of all prediction length (stated in Table 2) and the relative MSE reduction ratios(Promotion) by our method. Complete results can be found in Appendix B

表5：通过将我们提出的方法应用于变换器和RNN实现的性能提升。我们报告了所有预测长度（如表2所述）的平均MSE/MAE以及我们的方法实现的相对MSE降低比率（提升）。完整结果可以在附录B中找到。

| Dataset Model | Exchange | | ILI | | ETTm2 | | Electricity | | Traffic | | Weather | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| LSTM | 2.104 | 1.221 | 6.537 | 1.828 | 2.394 | 1.177 | 0.559 | 0.549 | 1.010 | 0.541 | 0.443 | 0.453 |
| +Ours | 1.294 | 0.946 | 4.305 | 1.442 | 1.338 | 0.896 | 0.381 | 0.437 | 0.755 | 0.430 | 0.277 | 0.333 |
| promotion | 38.49% | | 34.14% | | 44.11% | | 31.84% | | 25.24% | | 37.47% | |
| Transformer | 1.556 | 0.969 | 4.774 | 0.445 | 1.344 | 0.814 | 0.272 | 0.367 | 0.667 | 0.363 | 0.681 | 0.576 |
| +Ours | 1.271 | 0.874 | 4.471 | 1.394 | 1.254 | 0.806 | 0.256 | 0.364 | 0.662 | 0.359 | 0.615 | 0.537 |
| promotion | 18.31% | | 6.77% | | 6.69% | | 5.88% | | 0.75% | | 9.69% | |
| Informer | 1.550 | 0.998 | 5.136 | 1.544 | 1.410 | 0.822 | 0.31 | 0.396 | 0.764 | 0.415 | 0.633 | 0.548 |
| +Ours | 1.433 | 0.949 | 4.676 | 1.453 | 1.249 | 0.794 | 0.288 | 0.38 | 0.736 | 0.399 | 0.576 | 0.511 |
| promotion | 7.54% | | 8.95% | | 11.41% | | 7.09% | |  | 3.66% | 10.42% | |
| Autoformer | 0.613 | 0.539 | 3.006 | 1.161 | 0.324 | 0.367 | 0.227 | 0.337 | 0.627 | 0.378 | 0.337 | 0.381 |
| +Ours | 0.504 | 0.499 | 2.738 | 1.108 | 0.315 | 0.359 | 0.217 | 0.326 | 0.616 | 0.367 | 0.318 | 0.368 |
| promotion | 17.78% | | 8.91% | | 2.77% | | 4.40% | | 1.75% | | 5.63% | |
| Reformer | 1.620 | 1.023 | 4.724 | 1.445 | 1.479 | 0.915 | 0.337 | 0.422 | 0.740 | 0.421 | 0.802 | 0.655 |
| +Ours | 1.275 | 0.907 | 4.398 | 1.378 | 1.443 | 0.897 | 0.318 | 0.397 | 0.711 | 0.394 | 0.585 | 0.551 |
| promotion | 21.29% | | 6.90% | | 2.43% | | 5.63% | | 3.91% | | 27.05% | |

By analyzing the results of Table 5, we can obviously find that the module gains of the FECAM are large in datasets Exchange, ETTm2, and weather, but small in the dataset traffic. By observing the frequency spectrum of each dataset (as shown in the Fig. 2), we can safely say that datasets like Exchange, ETTm2, and Weather have a lot of energy information at low frequencies, while there is little energy information in the frequency spectrum of the Traffic dataset, ant this might be the reason why the gains of FECAM module is not so profitable on the traffic dataset.

通过分析表5的结果，我们可以明显地发现FECAM模块在Exchange、ETTm2和weather数据集上的增益较大，而在traffic数据集上的增益较小。通过观察每个数据集的频率谱（如图2所示），我们可以有把握地说，像Exchange、ETTm2和Weather这样的数据集在低频处有很多能量信息，而Traffic数据集的频率谱中能量信息较少，这可能是FECAM模块在traffic数据集上收益不高的原因。

# 4.2 Model Analysis

# 4.2 模型分析

Qualitative results: As shown in Fig.8, we plot the prediction results of vanilla Transformer, Transformer with our FECAM block, and Our FECAM method (FECAM with a projection layer) on Exchange dataset, and plot the prediction results of vanilla LSTM, LSTM with

定性结果：如图8所示，我们在Exchange数据集上绘制了vanilla Transformer、带有我们FECAM模块的Transformer以及我们的FECAM方法（带有投影层的FECAM）的预测结果，并在ETTm2数据集上绘制了vanilla LSTM、带有

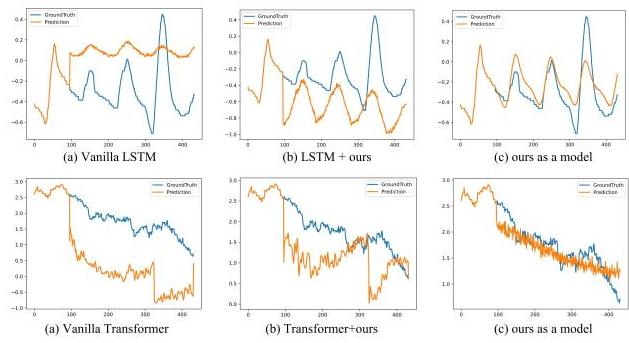


Figure 8: Visualization of ETTm2 and Exchange predictions given by different models.

图8：不同模型给出的ETTm2和Exchange预测的可视化。

FECAM block, and our FECAM method on ETTm2 dataset. When the input length is 96 steps and the output horizon is 336 steps, Transformer and LSTM both fail to capture the scale and bias of the future data on Exchange and ETTm2 respectively (as shown in Fig.8(a, d)) Moreover, transformer can hardly predict a proper trend on aperiodic data such as Exchange-Rate. With Our FECAM module, both Transformer and LSTM have a better predictability compared to their vanilla version (as shown in Fig.8(b, e)). We can see FECAM with just a projection layer can have a remarkable prediction result than other models with FECAM module, it could be the reason that parameters of FECAM is much less small than Transformer and LSTM which means Transformer and LSTM are more likely to get overfitting than FECAM with just a projection layer. These phenomena further indicate the inadequacy of existing mainstream models modelling in frequency for the TSF task. Our proposed method is beneficial for an accurate prediction of the detailed series variation, which is vital in real-world time series forecasting.

FECAM模块以及我们的FECAM方法在ETTm2数据集上的表现。当输入长度为96步，输出范围为336步时，Transformer和LSTM均未能捕捉到Exchange和ETTm2未来数据的规模和偏置（如图8(a, d)所示）。此外，Transformer在处理非周期性数据（如汇率）时几乎无法预测出适当趋势。使用我们的FECAM模块，与它们的原始版本相比，Transformer和LSTM的预测性能都有所提高（如图8(b, e)所示）。我们可以看到，仅带有投影层的FECAM与其他带有FECAM模块的模型相比，可以取得更显著的预测结果，这可能是因为FECAM的参数比Transformer和LSTM小得多，这意味着Transformer和LSTM比仅带有投影层的FECAM更容易过拟合。这些现象进一步表明，现有的主流模型在时间序列预测任务中对频率建模的不足。我们提出的方法对于准确预测详细序列变化是有益的，这对于实际时间序列预测至关重要。

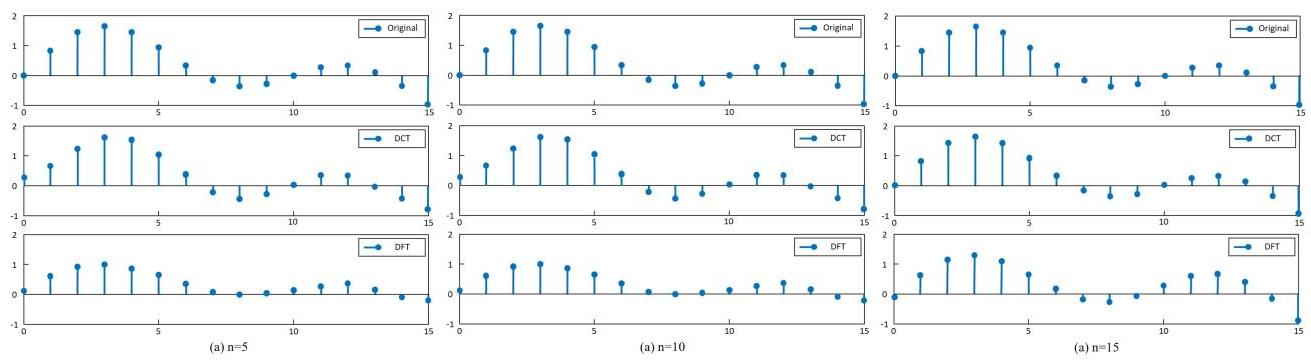


Figure 9: Signal reconstruction contrast between DCT and DFT with different number of frequency components.

图9：不同频率成分数量的DCT与DFT信号重建对比。

# 4.3 The interpretability of FECAM

# 4.3 FECAM的可解释性

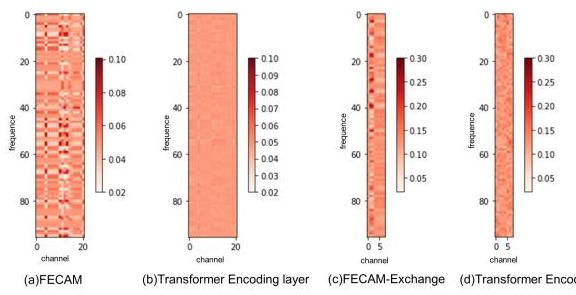


Figure 10: Visualization of frequency enhanced channel attention and output tensor of encoder layer of transformer.x-axis represents channels, y-axis represents frequency from low to high, performing on datasets weather and exchange.

图10：在weather和exchange数据集上，频率增强通道注意力和变换器编码层输出张量的可视化。x轴代表通道，y轴代表从低到高的频率。

Fig. 10(a) and (b) visualize the tensor of channel attention in FECAM and encoder layer of Transformer on the weather dataset, and Fig. 10(c) and (d) visualize the tensor of channel attention in FECAM and encoder layer of Transformer on the exchange dataset. we can see FECAM can extract the importance of different channels and the significance of different frequencies with obvious patterns compared with output tensor of encoder layer in transformer.

图10(a)和(b)可视化了FECAM和Transformer编码层在天气数据集上的通道注意力张量，图10(c)和(d)可视化了FECAM和Transformer编码层在交易所数据集上的通道注意力张量。我们可以看到，与Transformer编码层的输出张量相比，FECAM能够以明显的模式提取不同通道的重要性和不同频率的重要性。

# 4.4 Energy Compaction

# 4.4 能量压缩

For a signal with 16 sampling points, DCT and DFT are respectively used for reconstruction. During reconstruction, DCT and DFT use number of components starting from low frequency. The effect is as shown in Fig. 9.

对于一个具有16个采样点的信号，分别使用DCT和DFT进行重建。在重建过程中，DCT和DFT分别从低频开始使用 个组件。效果如图9所示。

This experiment intuitively verified that for a signal with more energy concentrated in the low frequency, DCT can better reconstruct the signal with using less components.So Discrete Cosine Transform is more efficient in Energy compaction than Discrete Fourier Transform.

这个实验直观地验证了，对于能量主要集中在低频的信号，DCT可以使用更少的组件更好地重建信号。因此，离散余弦变换在能量压缩方面比离散傅里叶变换更有效。

# 5 CONCLUSION

# 5 结论

This paper addresses time series forecasting from the perspective of modelling in frequency domain. Unlike previous studies that most frequency extraction method are FT-based which could bring high-frequency noise to the results due to the problematic periodity, which is known as Gibbs Phenomenon. We propose Frequency enhanced channel mechanism based on Discrete Cosine Transform could intrinsically avoid G-phenomenon, and we theoretically prove the feasibility of the method. By modeling in the frequency domain, FECAM can assign channel weights to different channels, and learn the importance of different frequencies of each channel, so as to learn the frequency domain representation of time series. In the experimental stage, we visualize the frequency domain information extracted by FECAM, which verify our conjecture and proved its validity. Most importantly, for its generalization, we design this method into a module for accessibility, which can flexibly and easily use in other mainstream model like transformer-based methods and RNNs methods, etc., with just a few lines to add. Our work achieve state-of-the-art on six real-world benchmarks. This impressive generality and performance of proposed frequency enhanced channel attention mechanism can be interesting of future research for time series forecasting.

本文从频域建模的角度探讨时间序列预测问题。与之前大多数研究采用基于傅里叶变换的频率提取方法不同，这些方法可能会因为周期性问题而给结果带来高频噪声，这被称为吉布斯现象。我们提出基于离散余弦变换的频率增强通道机制，能够本质上避免吉布斯现象，并且我们从理论上证明了该方法的可行性。通过在频域建模，FECAM可以为不同的通道分配权重，并学习每个通道不同频率的重要性，从而学习时间序列的频域表示。在实验阶段，我们可视化FECAM提取的频域信息，验证了我们的猜想并证明了其有效性。最重要的是，为了其泛化性，我们将该方法设计成一个可访问的模块，可以灵活且轻松地嵌入到其他主流模型中，如基于变换器的模型和RNN等方法，只需添加几行代码即可。我们的工作在六个真实世界基准测试中取得了最先进的结果。所提出的频率增强通道注意力机制的卓越泛化性和性能，对于时间序列预测的未来研究可能是有趣的。

# ACKNOWLEDGMENTS

# 致谢

This work was supported by the National Key R&D Program of China (No. 2020YFB1709702) and the National Natural Science Foundation of China (No. 62073313).

本工作得到了中国国家重点研发计划（编号：2020YFB1709702）和中国国家自然科学基金（编号：62073313）的支持。

# REFERENCES

# 参考文献

[1] O. Anderson and M. Kendall. 1976. Time-Series. 2nd edn. J. R. Stat. Soc. (Series D) (1976).

[2] Silvio Barra, Salvatore Mario Carta, Andrea Corriga, Alessandro Sebastian Podda, encoding for financial forecasting. IEEE/CAA Journal of Automatica Sinica 7, 3 (2020), 683-692.

[3] Jian Cao, Zhi Li, and Jian Li. 2019. Financial time series forecasting model based on CEEMDAN and LSTM. Physica A: Statistical mechanics and its applications 519 (2019), 127-139.

[4] Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza, Max Mergen-thaler, and Artur Dubrawski. 2022. N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting. arXiv preprint arXiv:2201.12886 (2022). Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision Transformer: Reinforcement Learning via Sequence Modeling. NeurIPS (2021).

[6] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT.

[7] Qianggang Ding, Sifan Wu, Hao Sun, Jiadong Guo, and Jian Guo. 2020. Hierarchical Multi-Scale Gaussian Transformer for Stock Movement Prediction.. In IJCAI. 4640-4646.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In ICLR. https://openreview.net/forum?id=YicbFdNTTy

[9] J Foster and FB Richards. 1991. The Gibbs phenomenon for piecewise-linear approximation. The American Mathematical Monthly 98, 1 (1991), 47-49.

[10] Marzyeh Ghassemi, Marco Pimentel, Tristan Naumann, Thomas Brennan, David Clifton, Peter Szolovits, and Mengling Feng. 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In Proceedings of the AAAI conference on artificial intelligence, Vol. 29.

[11] Tilmann Gneiting and Adrian E Raftery. 2005. Weather forecasting with ensemble methods. Science 310, 5746 (2005), 248-249.

[12] Klaus Greff, Rupesh K Srivastava, Jan Koutnik, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems 28, 10 (2016), 2222-2232.

[13] Aditya Grover, Ashish Kapoor, and Eric Horvitz. 2015. A deep hybrid model for weather forecasting. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 379-386.

[14] Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. 2021. Multiwavelet-based Operator Learning for Differential Equations. arXiv:2109.13459 [cs.LG]

[15] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. IEEE Conference on Computer Vision and Pattern Recognition.

[16] Sasikiran Kandula, Teresa Yamana, Sen Pei, Wan Yang, Haruka Morita, and Jeffrey Shaman. 2018. Evaluation of mechanistic and statistical methods in forecasting influenza-like illness. Journal of The Royal Society Interface 15, 144 (2018), 20180174.

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In ICLR. http://arxiv.org/abs/1412.6980

[18] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. In ICLR. https://openreview.net/forum?id=rkgNKkHtvB

[19] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In SIGIR. and Xifeng Yan. 2019. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In NeurIPS. https://proceedings. neurips.cc/paper/2019/file/6775a0635c302542da2c32aa19d86be0-Paper.pdf

[21] Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. 2022. SCINet: Time Series Modeling and Forecasting with Sample Convolution and Interaction. Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS), 2022 (2022).

[22] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and long-range time series modeling and forecasting. In ICLR.

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In ICCV.

[24] Tao Ma, Constantinos Antoniou, and Tomer Toledo. 2020. Hybrid machine learning algorithm and statistical time series model for network-wide traffic forecast. Transportation Research Part C: Emerging Technologies 111 (2020), 352- 372.

[25] Danielle C Maddix, Yuyang Wang, and Alex Smola. 2018. Deep factors with gaussian processes for forecasting. arXiv preprint arXiv:1812.00098 (2018).

[26] E Moskona, P Petrushev, and EB Saff. 1995. The Gibbs phenomenon for bestL 1-trigonometric polynomial approximation. Constructive Approximation 11, 3

[27] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. ICLR (2019).

[28] Adam Paszke, S. Gross, Francisco Massa, A. Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Z. Lin, N. Gimelshein, L. Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In NeurIPS.

[29] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. 2018. Deep state space models for time series forecasting. In NeurIPS.

[30] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. 2020. WeatherBench: a benchmark data set for data-driven weather forecasting. Journal of Advances in Modeling Earth Systems 12, 11 (2020), e2020MS002203.

[31] Andrew G Reece, Andrew J Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M Danforth, and Ellen J Langer. 2017. Forecasting the onset and course of mental illness with Twitter data. Scientific reports 7, 1 (2017), 1-11.

[32] Luis G Baca Ruiz, R Rueda, Manuel P Cuéllar, and MC Pegalajar. 2018. Energy consumption forecasting based on Elman neural networks with evolutive optimization. Expert Systems with Applications 92 (2018), 380-389.

[33] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. Int. J.

[34] Bernie D Shizgal and Jae-Hun Jung. 2003. Towards the resolution of the Gibbs phenomena. J. Comput. Appl. Math. 161, 1 (2003), 41-65. https://doi.org/10.1016/ S0377-0427(03)00500-4

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In NeurIPS. https://proceedings.neurips.cc/paper/2017/file/ 3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[36] Huaizhi Wang, Zhenxing Lei, Xian Zhang, Bin Zhou, and Jianchun Peng. 2019. and Management 198 (2019), 111799.

[37] Nan Wei, Changjun Li, Xiaolong Peng, Fanhua Zeng, and Xinqian Lu. 2019. Conventional models and artificial intelligence-based models for energy consumption forecasting: A review. Journal of Petroleum Science and Engineering 181 (2019), 106187.

[38] Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. 2017. A multi-horizon quantile recurrent forecaster. NeurIPS (2017).

[39] Steven Wong, Robin Walters, Lejun Jiang, Tamas G Molnar, and Rose Yu. 2020. ral Networks. NIPS., Dec (2020).

[40] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In NeurIPS.

[41] Yinjun Wu, Jingchao Ni, Wei Cheng, Bo Zong, Dongjin Song, Zhengzhang Chen, Yanchi Liu, Xuchao Zhang, Haifeng Chen, and Susan B Davidson. 2021. Dynamic Gaussian mixture based deep generative model for robust forecasting on sparse multivariate time series. In Proceedings of the AAAI Conference on Artificial

[42] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. 2020. Learning in the Frequency Domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[43] Rose Yu, Stephan Zheng, Anima Anandkumar, and Yisong Yue. 2017. Long-term forecasting using tensor-train RNNs. arXiv preprint arXiv:1711.00073 (2017).

[44] Pengyu Zeng, Guoliang Hu, Xiaofeng Zhou, Shuai Li, Pengjie Liu, and Shurui Liu. 2022. Muformer: A long sequence time-series forecasting model based on modified multi-head attention. Knowledge-Based Systems 254 (2022), 109584.

[45] Chaoyun Zhang and Paul Patras. 2018. Long-term mobile traffic forecasting using deep spatio-temporal neural networks. In Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing. 231-240.

[46] Qi Zhang, Jianlong Chang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2020. Spatio-temporal graph structure learning for traffic forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 1177-1185.

[47] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In AAAI.