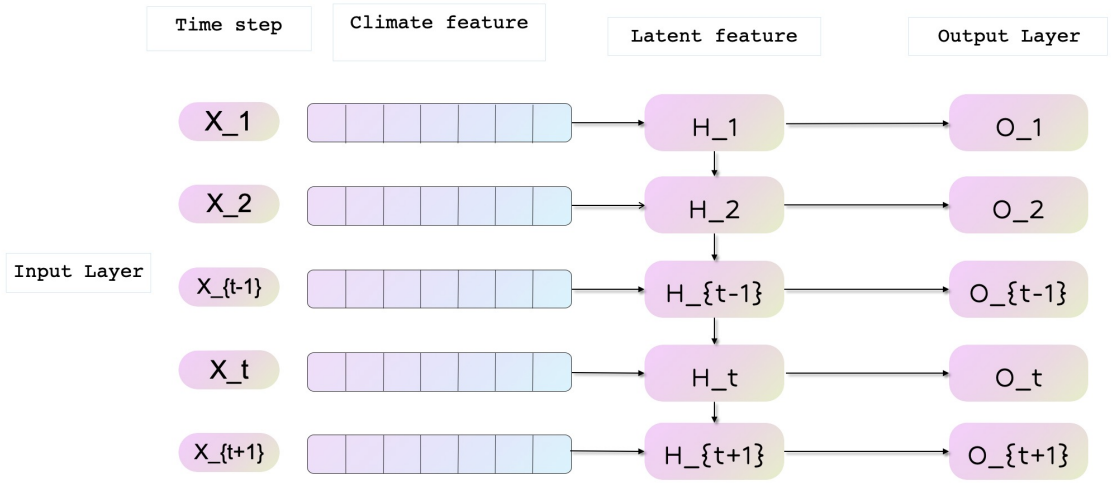
Week 3

Wentao Gao

RNN Architecture



$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

$$y_t = W_{hy}h_t + b_y$$

LSTM Architecture

Forget gate: $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$

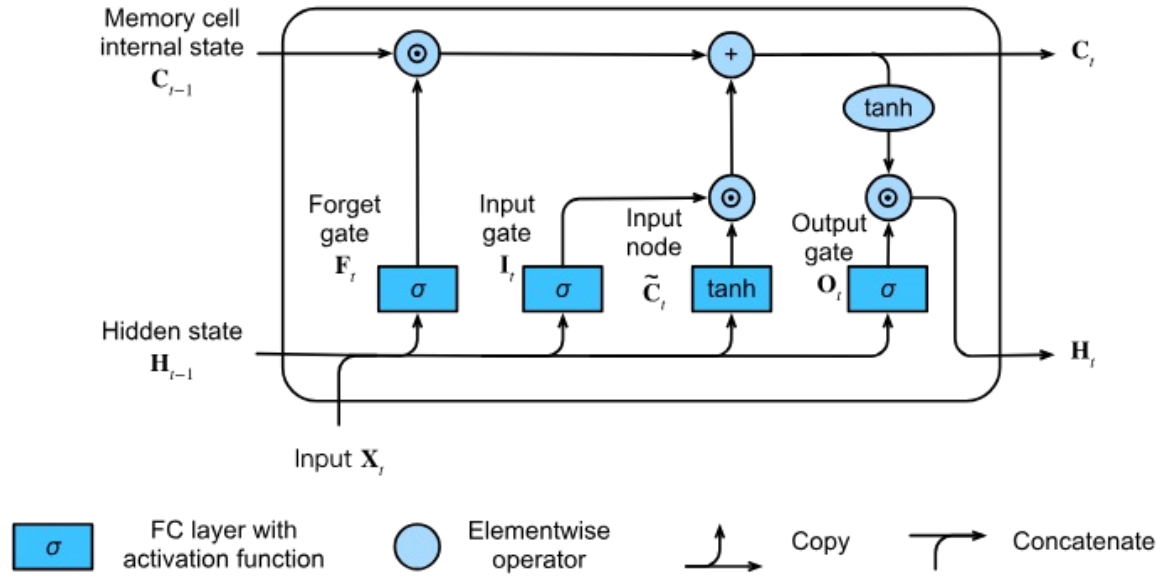
Input gate: $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$

Candidate cell state: $\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C)$

Update cell state: $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$

Output gate: $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$

Hidden state: $h_t = o_t \odot \tanh(C_t)$



Transformer Architecture

- Scaled Dot-Product Attention:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Multi-Head Attention:

$$MultiHead(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O$$

- Position-wise Feed-Forward Networks:

$$\text{head}_i = \text{Attention}(QW_{Qi}, KW_{Ki}, VW_{Vi})$$

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

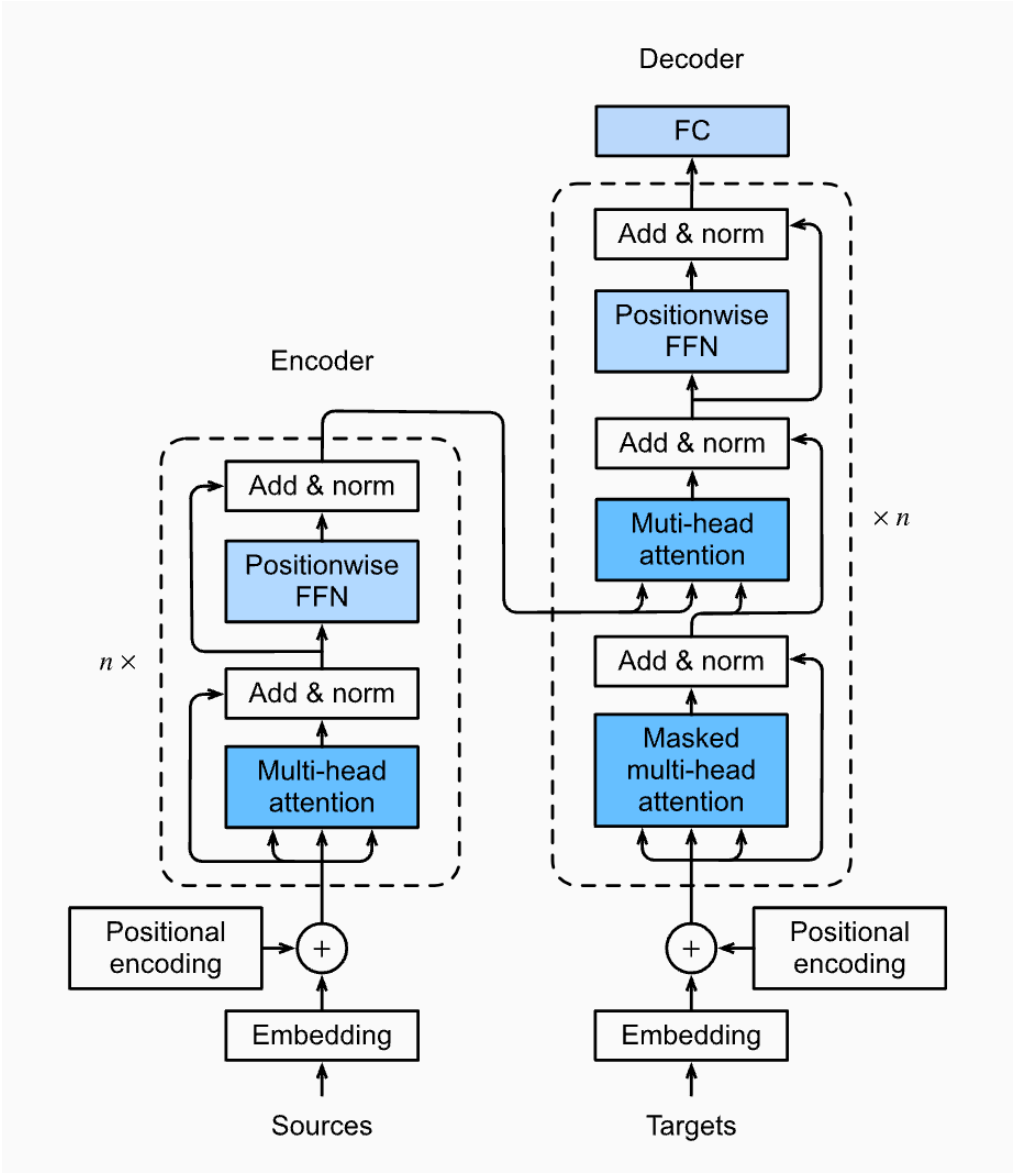
- Positional Encoding:

$$PE_{(\text{pos},2i)} = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right)$$

$$PE_{(\text{pos},2i+1)} = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right)$$

- Layer Normalization:

$$y = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}$$



Transformer Block

A **transformer block** is a parameterized function class $f_\theta : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$. If $\mathbf{x} \in \mathbb{R}^{n\times d}$ then $f_\theta(\mathbf{x}) = \mathbf{z}$ where

$$Q^{(h)}(\mathbf{x}_i) = W_{h,q}^T \mathbf{x}_i, \quad K^{(h)}(\mathbf{x}_i) = W_{h,k}^T \mathbf{x}_i, \quad V^{(h)}(\mathbf{x}_i) = W_{h,v}^T \mathbf{x}_i, \qquad W_{h,q}, W_{h,k}, W_{h,v} \in \mathbb{R}^{d\times k}, \tag{1}$$

$$\alpha_{i,j}^{(h)} = \text{softmax}_j\left(\frac{\langle Q^{(h)}(\mathbf{x}_i), K^{(h)}(\mathbf{x}_j)\rangle}{\sqrt{k}}\right), \tag{2}$$

$$\mathbf{u}_i' = \sum_{h=1}^{H} W_{c,h}^T \sum_{j=1}^{n} \alpha_{i,j}^{(h)} V^{(h)}(\mathbf{x}_j), \qquad W_{c,h} \in \mathbb{R}^{k\times d}, \tag{3}$$

$$\mathbf{u}_i = \text{LayerNorm}(\mathbf{x}_i + \mathbf{u}_i'; \gamma_1, \beta_1), \qquad \gamma_1, \beta_1 \in \mathbb{R}^d, \tag{4}$$

$$\mathbf{z}_i' = W_2^T \text{ReLU}(W_1^T \mathbf{u}_i), \qquad W_1 \in \mathbb{R}^{d\times m}, W_2 \in \mathbb{R}^{m\times d}, \tag{5}$$

$$\mathbf{z}_i = \text{LayerNorm}(\mathbf{u}_i + \mathbf{z}_i'; \gamma_2, \beta_2), \qquad \gamma_2, \beta_2 \in \mathbb{R}^d. \tag{6}$$

The notation softmax_j indicates we take the softmax (defined in Equation 9) over the d-dimensional vector indexed by j. The LayerNorm function [Lei Ba et al., 2016] is defined for $\mathbf{z} \in \mathbb{R}^k$ by

$$\text{LayerNorm}(\mathbf{z}; \gamma, \beta) = \gamma\frac{(\mathbf{z} - \mu_\mathbf{z})}{\sigma_\mathbf{z}} + \beta, \qquad \gamma, \beta \in \mathbb{R}^k. \tag{7}$$

$$\mu_\mathbf{z} = \frac{1}{k}\sum_{i=1}^{k} \mathbf{z}_i, \quad \sigma_\mathbf{z} = \sqrt{\frac{1}{k}\sum_{i=1}^{k} (\mathbf{z}_i - \mu_\mathbf{z})^2}. \tag{8}$$

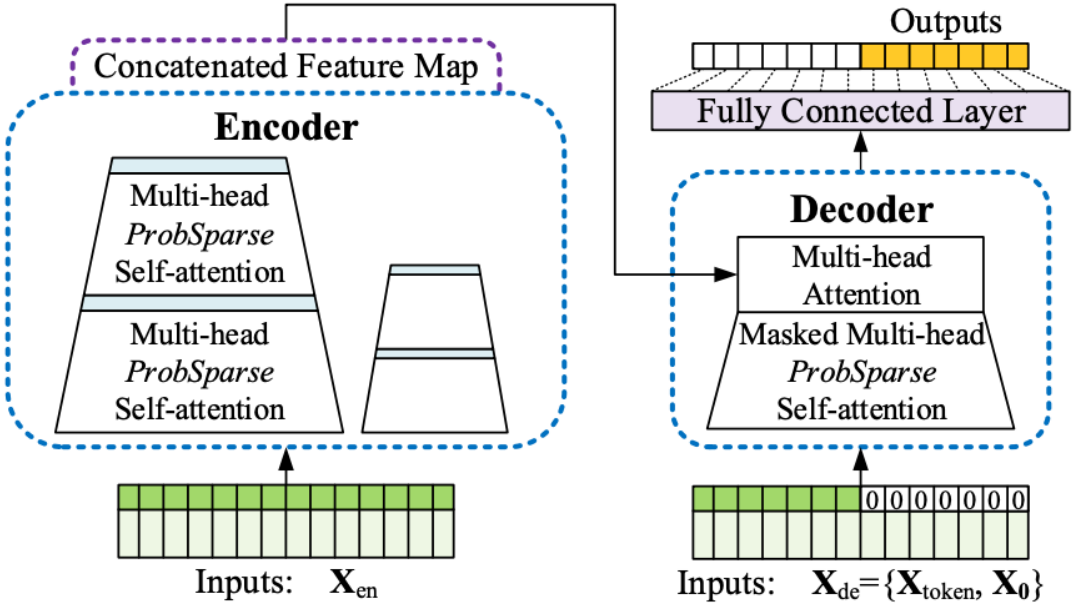
Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecastin	The Informer is an efficient deep learning model, designed specifically for handling large-scale, high-frequency, and long-range time series forecasting, and it introduces innovative techniques such as **ProbSparse self-attention, decomposed positional encoding, and generative-discriminative training** to improve the efficiency and predictive performance in long sequence handling.
Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting	To tackle the intricate temporal patterns of the long-term future, we present Autoformer as a ***decomposition architecture*** and design the inner decomposition block to empower the deep forecasting model with immanent progressive decomposition capacity. propose an ***Auto-Correlation mechanism*** with dependencies discovery and information aggregation at the series level. Our mechanism is beyond previous self-attention family and can simultaneously benefit the computation efficiency and information utilization.
PYRAFORMER: LOW-COMPLEXITY PYRAMIDAL ATTENTION FOR LONG-RANGE TIME SERIES MODELING AND FORECASTING	In this paper the authors introduce "**pyramidal attention module (PAM)** in which the inter-scale tree structure summarizes features at different resolutions and the intra-scale neighboring connections model the temporal dependencies of different ranges."
FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting (mlr.press)	For the long time series prediction problem, the authors propose a FEDformer model based on **frequency domain decomposition**. The prediction accuracy and model operation efficiency are greatly improved. The authors propose a **Fourier/wavelet transform-based module** that enables the model to achieve linear complexity while improving accuracy by performing a fixed number of random samples in the frequency domain.
Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting	This article investigates time series forecasting from the perspective of stationarity. We propose a method to enhance the **stationarity of the sequence** and improve the attention mechanism within the Transformer. This method reintroduces non-stationary information while enhancing the predictability of the data and unleashing the excellent temporal modeling capabilities of the attention mechanism. *Series Stationarization and De-stationary Attention*
LightTS: Less Is More: Fast Multivariate Time Series Forecasting with Light Sampling-oriented MLP Structures (arxiv.org)	The key idea of LightTS is to apply an **MLP-based structure** on top of two delicate down-sampling strategies, including **interval sampling** and **continuous sampling**, inspired by crucial characteristics of multivariate time series

FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting (mlr.press)	For the long time series prediction problem, the authors propose a FEDformer model based on **frequency domain decomposition**. The prediction accuracy and model operation efficiency are greatly improved. The authors propose a **Fourier/wavelet transform-based module** that enables the model to achieve linear complexity while improving accuracy by performing a fixed number of random samples in the frequency domain.
Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting	This article investigates time series forecasting from the perspective of stationarity. We propose a method to enhance the **stationarity of the sequence** and improve the attention mechanism within the Transformer. This method reintroduces non-stationary information while enhancing the predictability of the data and unleashing the excellent temporal modeling capabilities of the attention mechanism. *Series Stationarization and De-stationary Attention*
LightTS: Less Is More: Fast Multivariate Time Series Forecasting with Light Sampling-oriented MLP Structures (arxiv.org)	The key idea of LightTS is to apply an **MLP-based structure** on top of two delicate down-sampling strategies, including **interval sampling** and **continuous sampling**, inspired by crucial characteristics of multivariate time series
ETSformer: Exponential Smoothing Transformers for Time-series Forecasting	Brings the time-tested ideas of *seasonal-trend decomposition and exponential weighting* into the modern *transformers* framework. Seasonality and trend are critical components of time-series data, and ETSformer bakes these time-series priors into the architecture of a transformer model. Develop the **Exponential Smoothing Attention (ESA) and Frequency Attention (FA)** mechanisms to extract latent growth and seasonal representations
DLinear: Are Transformers Effective for Time Series Forecasting?	Doubt about transformer and Proposed a new baseline **DLinear**.
Long-term Forecasting with TiDE: Time-series Dense Encoder	
TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis	Focusing on the key problem of modeling temporal changes, this paper innovatively **transforms one-dimensional time series into two-dimensional space for analysis**, and further proposes a **task-general temporal basis model - TimesNet,** which achieves a comprehensive lead in five major mainstream temporal analysis tasks: long-time, short-time prediction, missing value filling, anomaly detection, and classification.
Fourier-Mixed Window Attention: Accelerating Informer for Long Sequence Time-Series Forecasting	propose to replace ProbSparse attention of Informer via a (local) window attention followed by a Fourier transform (mixing) layer, a novel local-global attention which we call Fourier-Mixed window attention (FWin).

Paper reading

Informer

- ***ProbSparse self-attention,***
- ***Decomposed positional encoding***
- ***Generative-discriminative training***



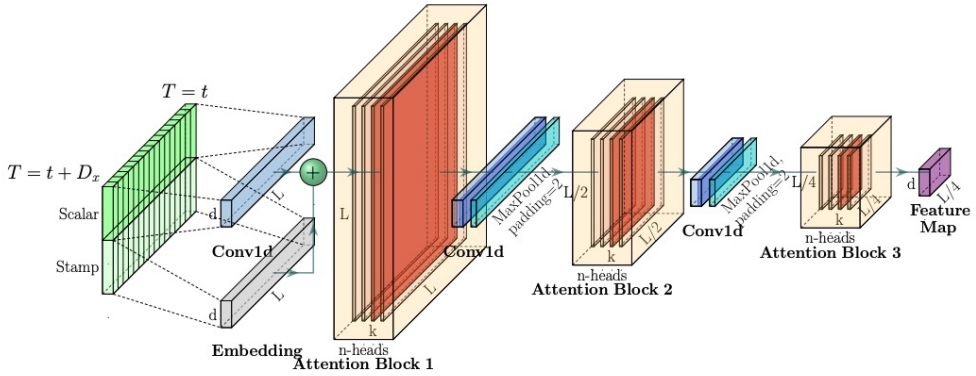
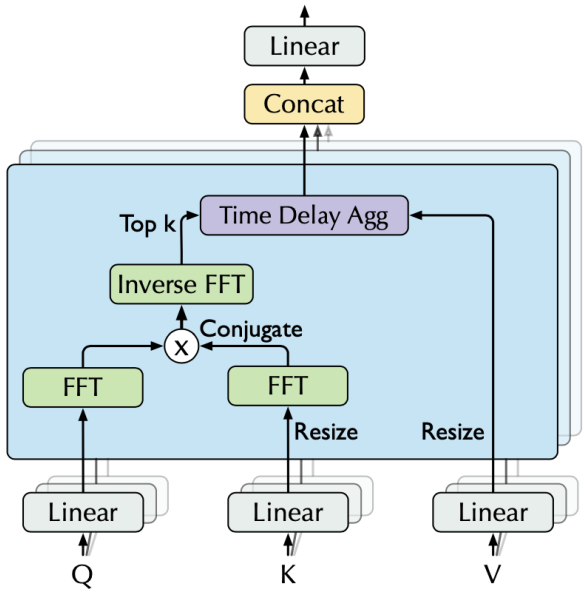
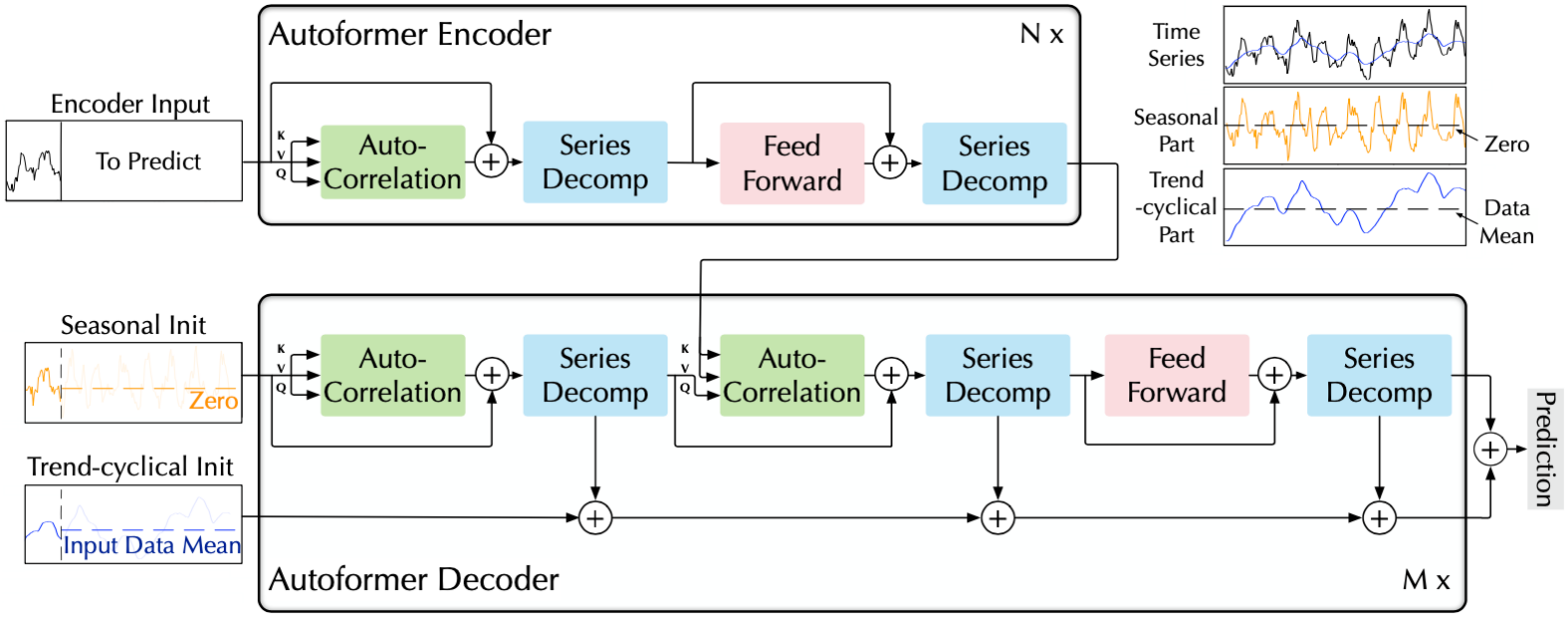


Figure 3: The single stack in Informer's encoder. (1) The horizontal stack stands for an individual one of the encoder replicas in Fig.(2). (2) The presented one is the main stack receiving the whole input sequence. Then the second stack takes half slices of the input, and the subsequent stacks repeat. (3) The red layers are dot-product matrixes, and they get cascade decrease by applying self-attention distilling on each layer. (4) Concatenate all stacks' feature maps as the encoder's output.

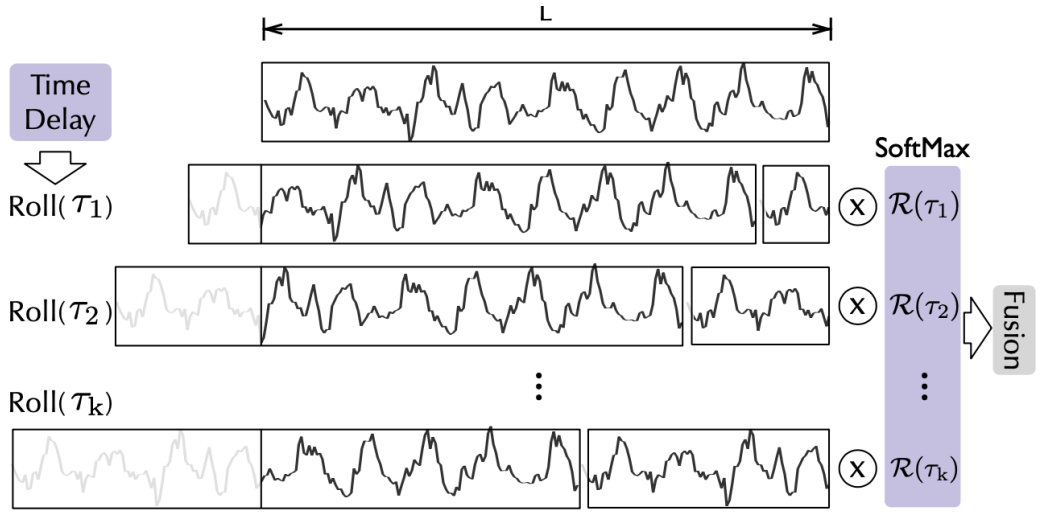




AutoFormer

Decomposition architecture

Auto-Correlation mechanism



FedFormer

Frequency domain decomposition

Fourier/wavelet transform-based module

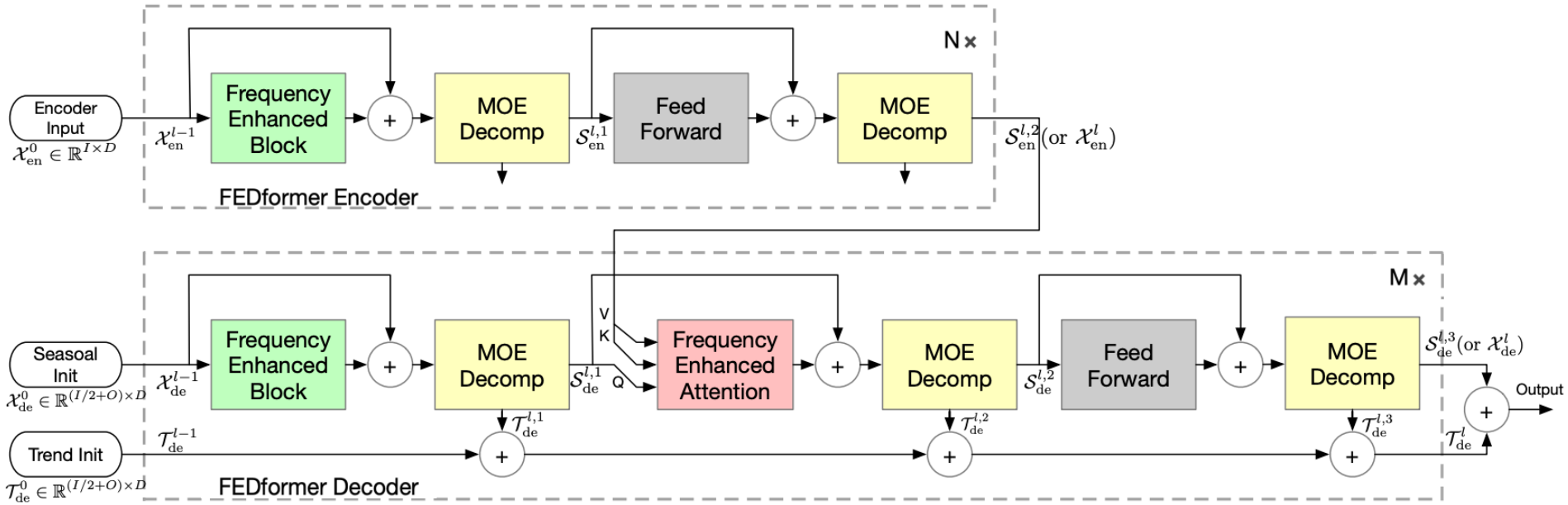


Figure 2. FEDformer Structure. The FEDformer consists of N encoders and M decoders. The Frequency Enhanced Block (FEB, green blocks) and Frequency Enhanced Attention (FEA, red blocks) are used to perform representation learning in frequency domain. Either FEB or FEA has two subversions (FEB-f & FEB-w or FEA-f & FEA-w), where '-f' means using Fourier basis and '-w' means using Wavelet basis. The Mixture Of Expert Decomposition Blocks (MOEDecomp, yellow blocks) are used to extract seasonal-trend patterns from the input data.

FedFormer

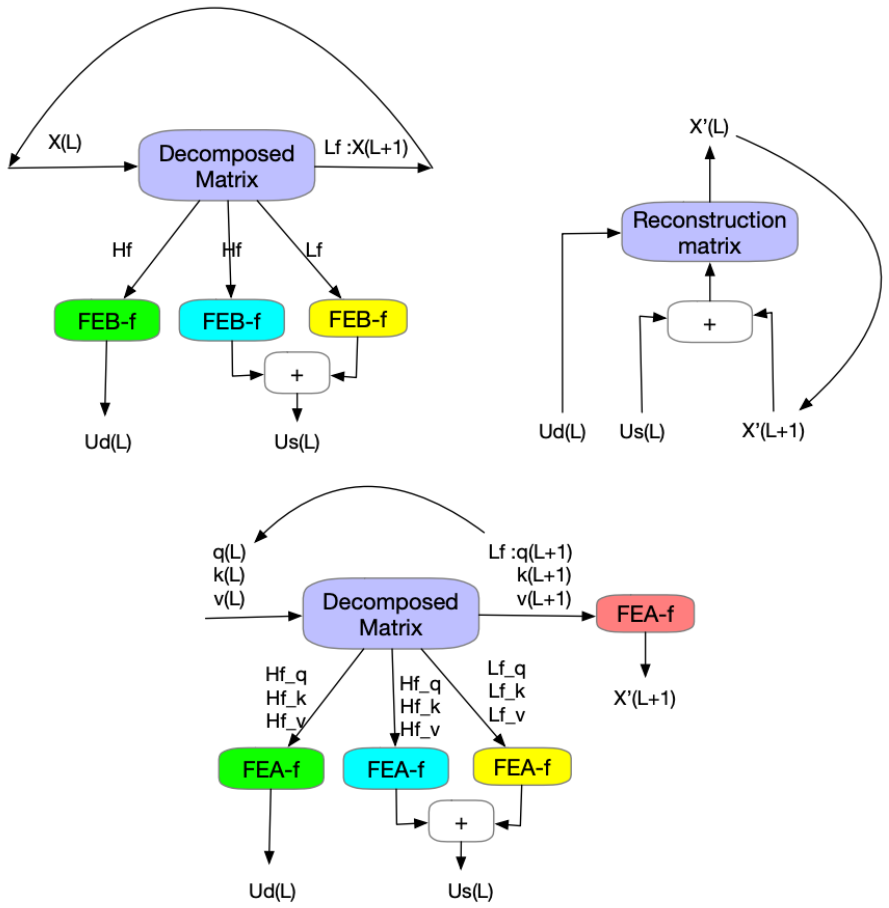


Figure 5. Top Left: Wavelet frequency enhanced block decomposition stage. Top Right: Wavelet block reconstruction stage shared by FEB-w and FEA-w. Bottom: Wavelet frequency enhanced cross attention decomposition stage.

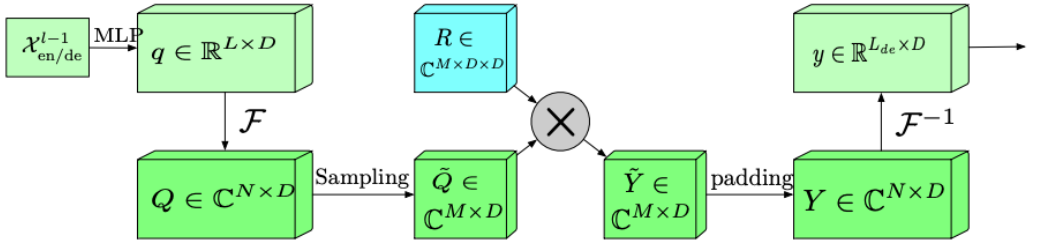


Figure 3. Frequency Enhanced Block with Fourier transform (FEB-f) structure.

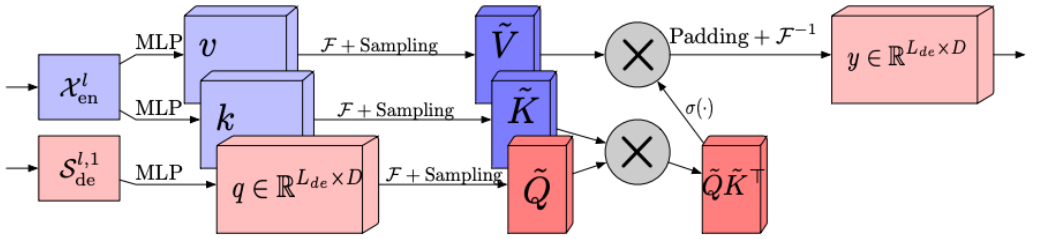
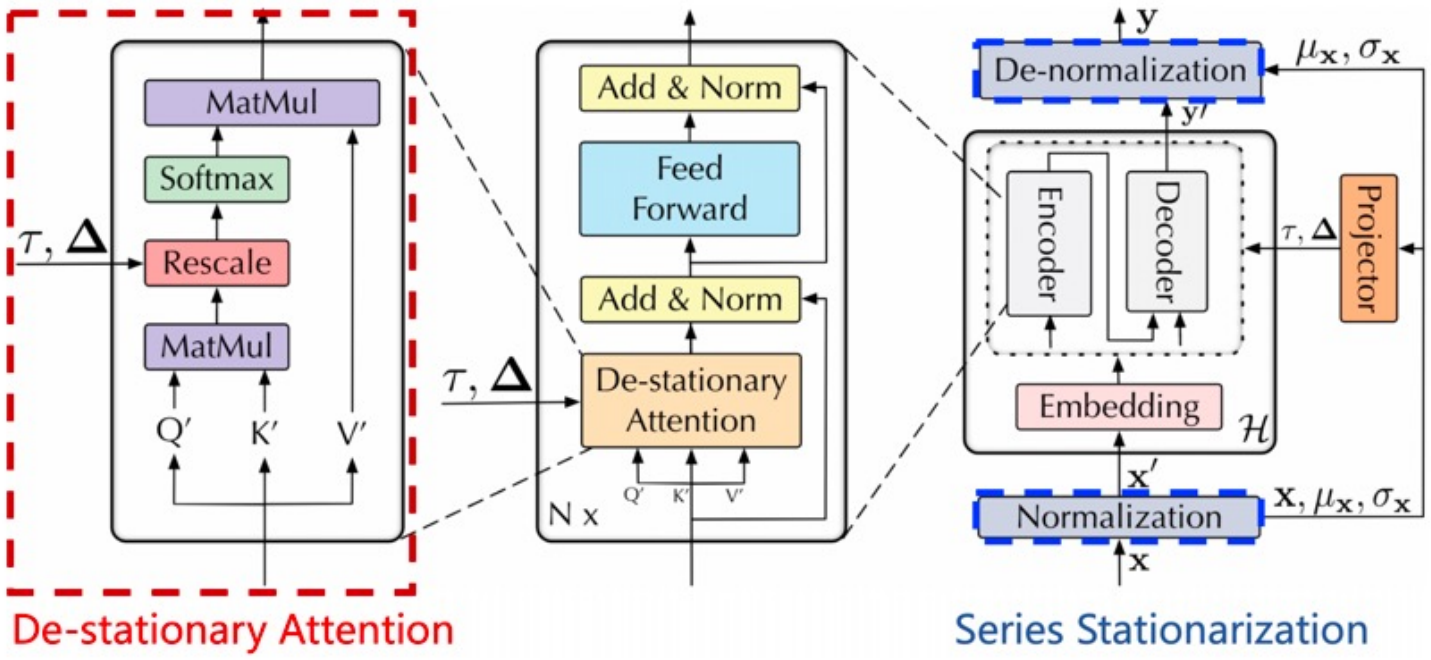
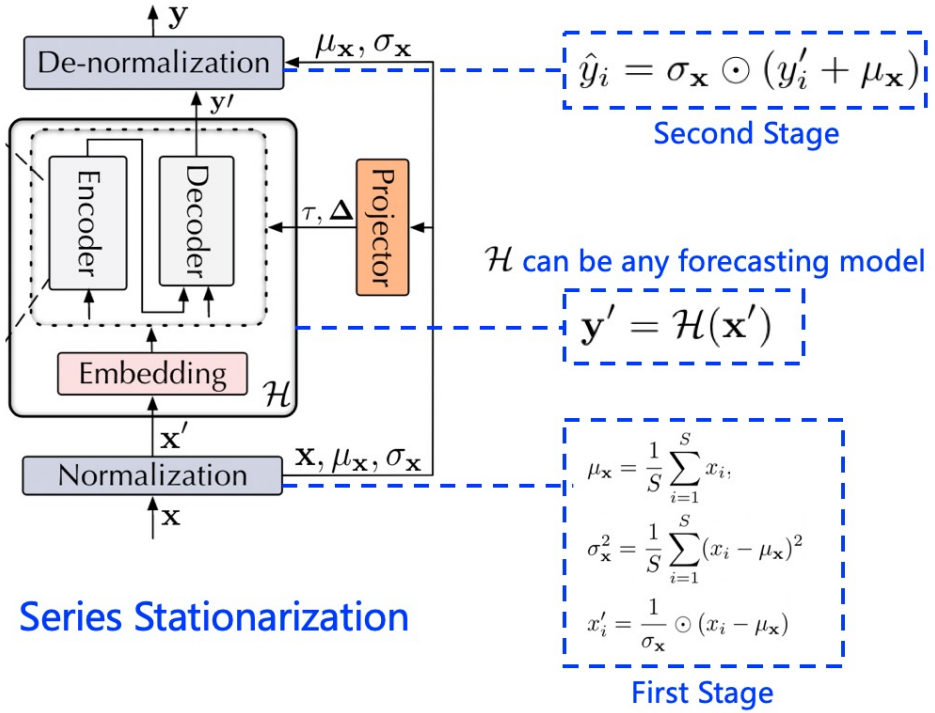


Figure 4. Frequency Enhanced Attention with Fourier transform (FEA-f) structure, $\sigma(\cdot)$ is the activation function.

Non-stationary Transformers





Series Stationarization and De-stationary Attention

LightTS

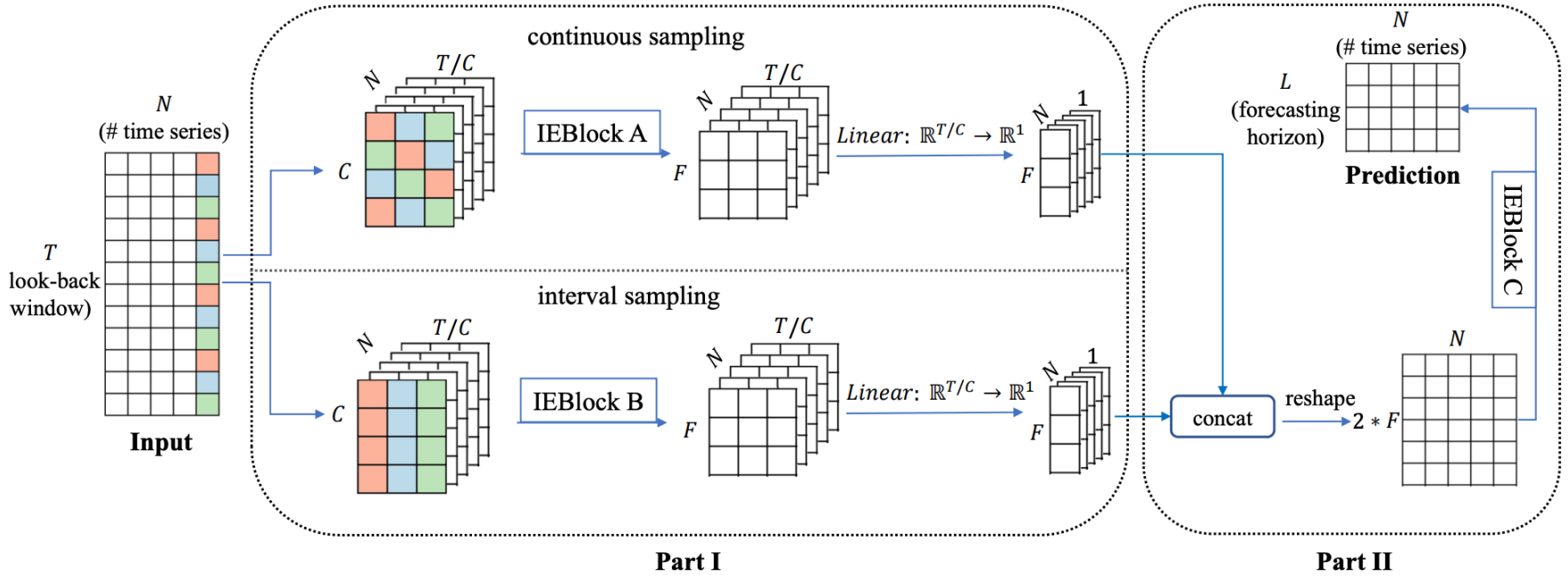


Figure 1: The overview of LightTS. In Part I, the model captures the short/long-term dependencies and extract features of each time series. In Part II, the model learns the interdependencies among different time series and make predictions.

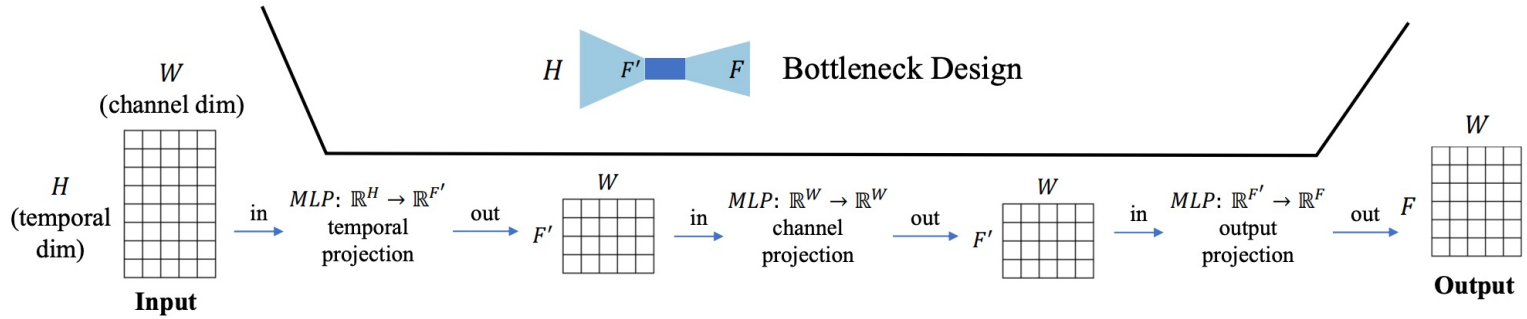


Figure 2: The overview of IEBlock and the bottleneck design.

MLP-based structure

interval sampling and continuous sampling

seasonal-trend decomposition and exponential weighting

ETSformer

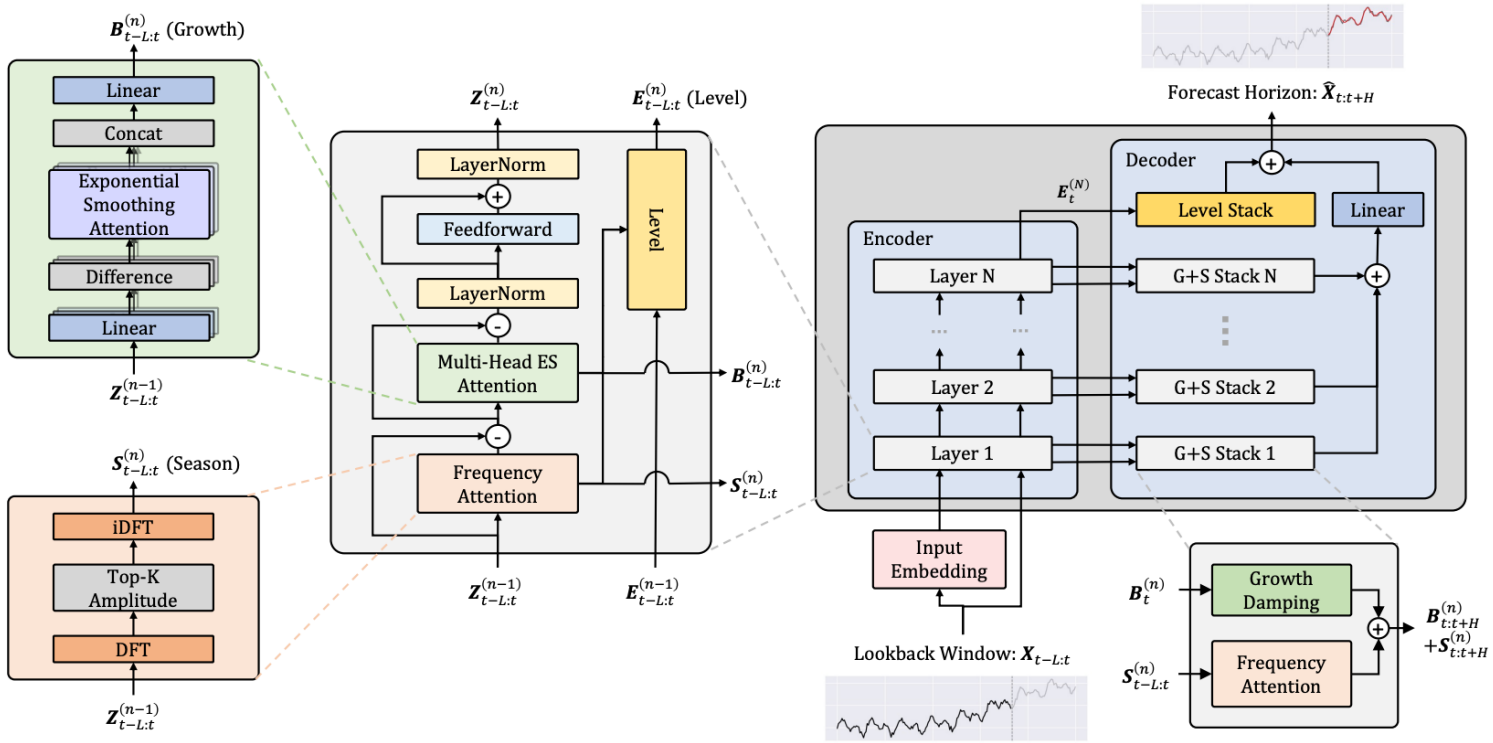
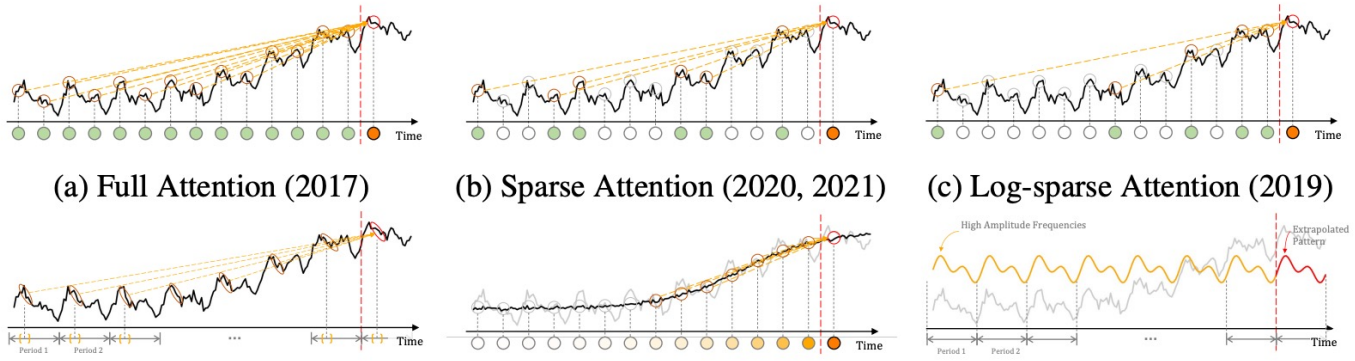


Figure 2: ETSformer model architecture.

Exponential Smoothing Attention (ESA) and Frequency Attention (FA)



(a) Full Attention (2017) (b) Sparse Attention (2020, 2021) (c) Log-sparse Attention (2019)

(d) Autocorrelation Attention (2021) (e) Exponential Smoothing Attention (Ours) (f) Frequency Attention (Ours)

Are Transformers Effective for Time Series Forecasting?

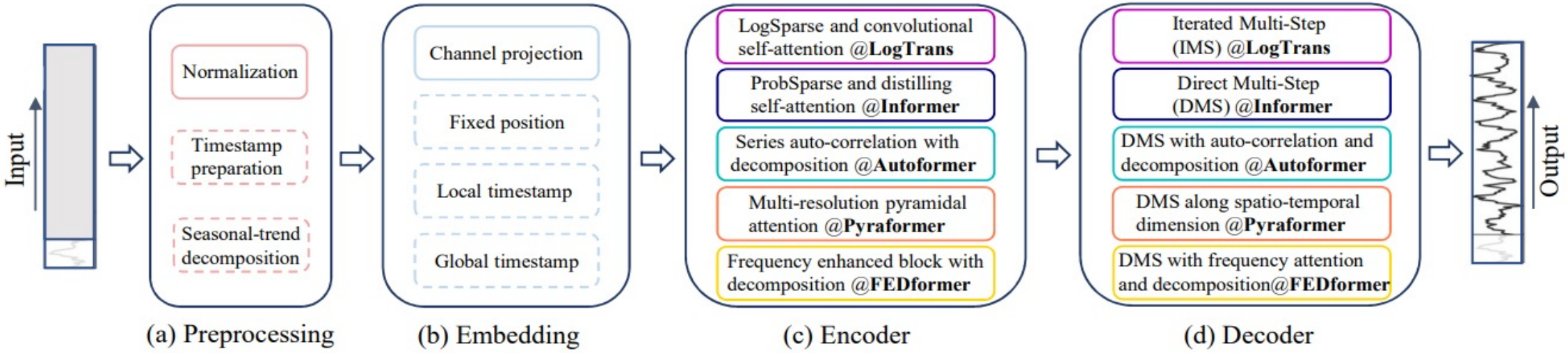
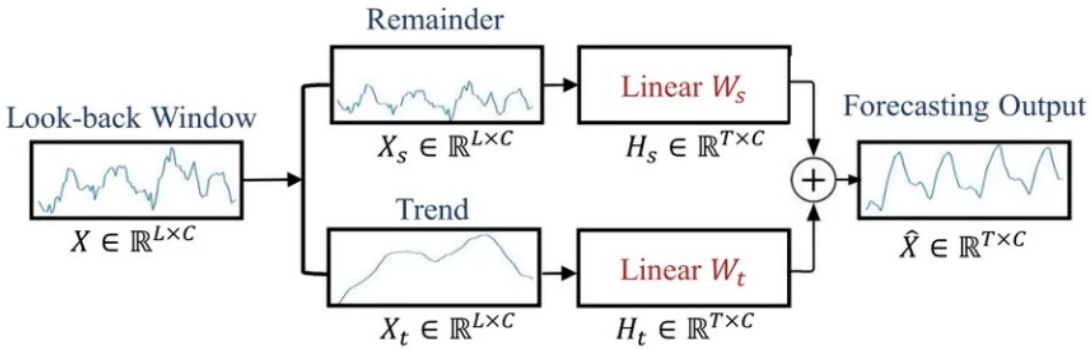


Figure 1. The pipeline of existing Transformer-based TSF solutions. In (a) and (b), the solid boxes are essential operations, and the dotted boxes are applied optionally. (c) and (d) are distinct for different methods [16, 18, 28, 30, 31].



(a) The whole structure of *DLinear*

DLinear

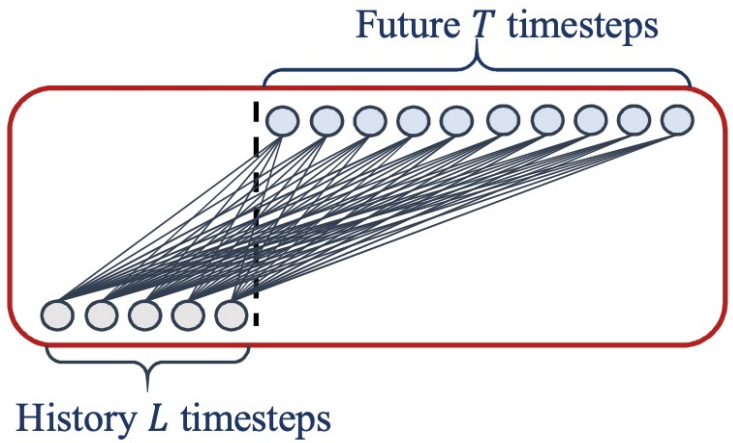


Figure 2: Illustration of the basic linear model.

TimesNet

- ***Transforms one-dimensional time series into two-dimensional space for analysis***
- ***Task-general temporal basis model - TimesNet***

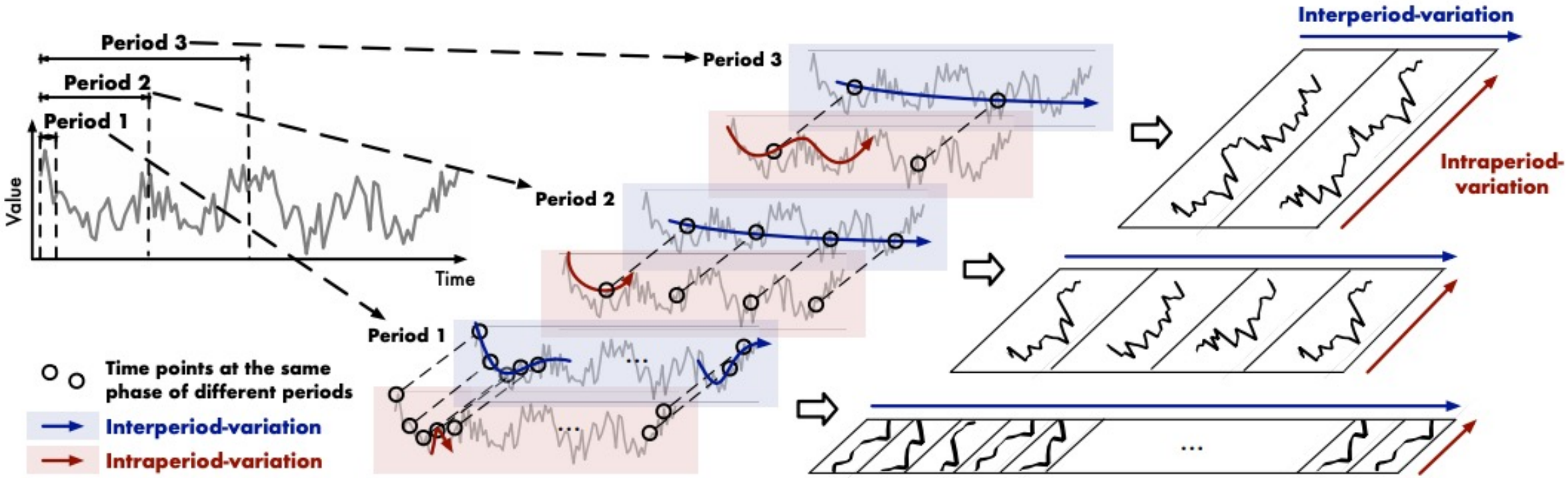
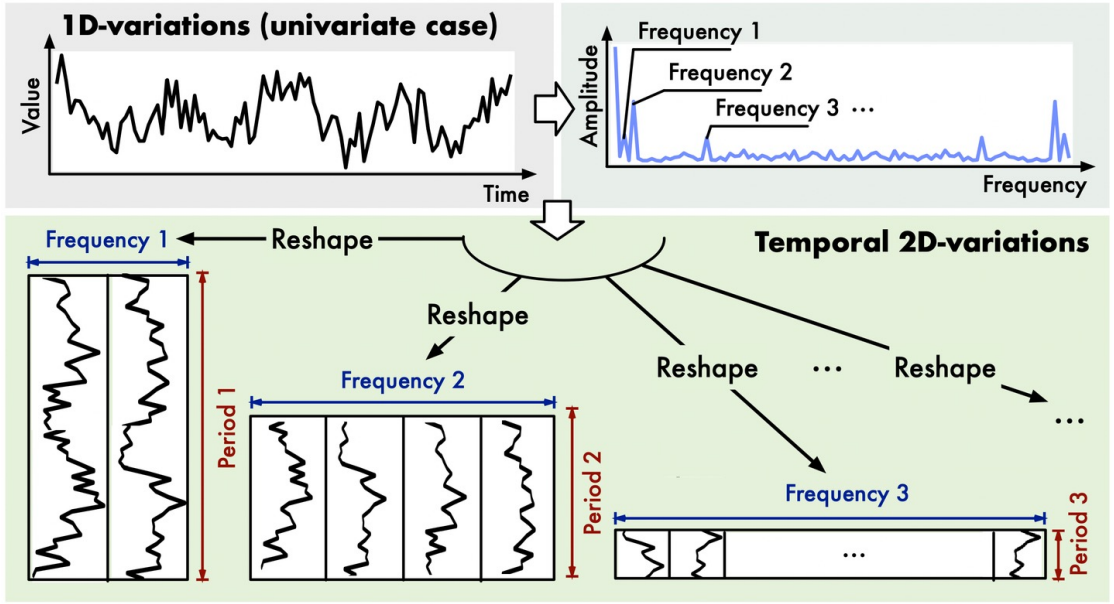
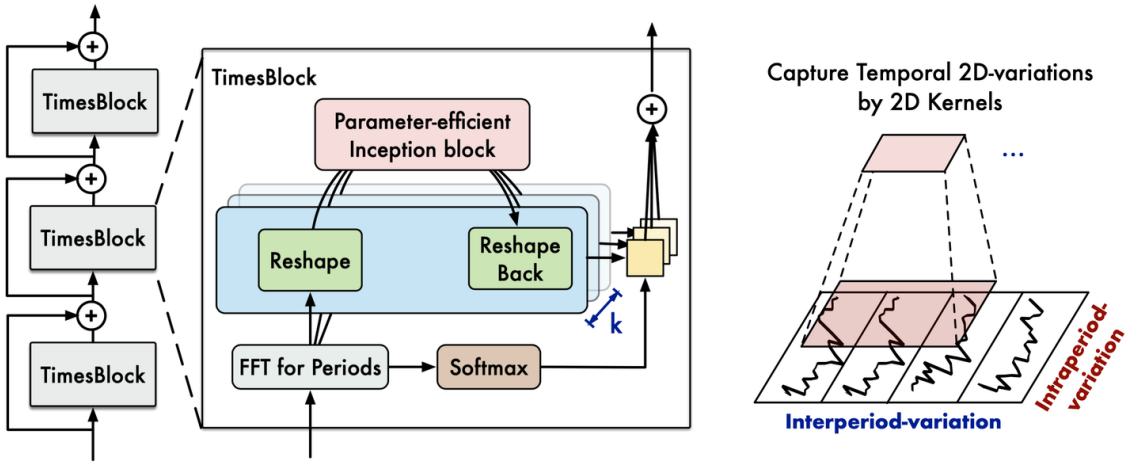
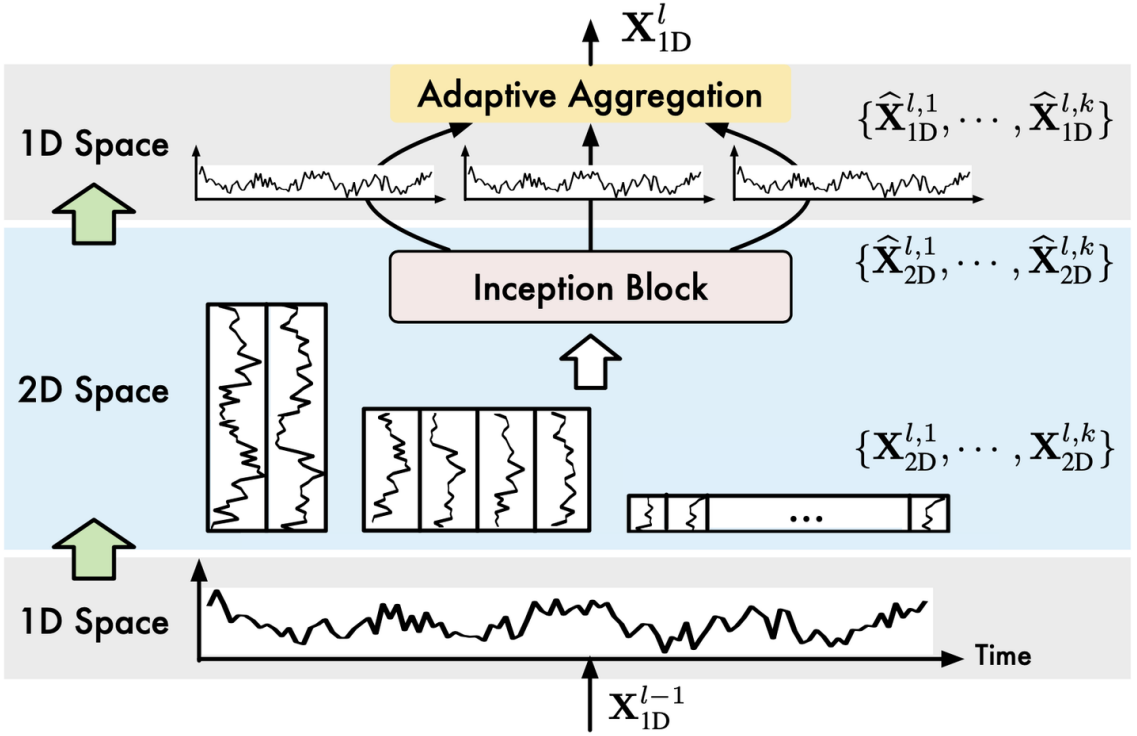


Figure 1: Multi-periodicity and temporal 2D-variation of time series. Each period involves the intraperiod-variation and interperiod-variation. We transform the original 1D time series into a set of 2D tensors based on multiple periods, which can unify the intraperiod- and interperiod-variations.


1D-variations (univariate case)
Value
Time
Amplitude
Frequency 1
Frequency 2
Frequency 3 ···
Frequency
Temporal 2D-variations
Reshape
Frequency 1
Frequency 2
Frequency 3
Period 1
Period 2
Period 3


TimesBlock
TimesBlock
TimesBlock
TimesBlock
Parameter-efficient Inception block
Reshape
Reshape Back
k
FFT for Periods
Softmax
Capture Temporal 2D-variations by 2D Kernels
Interperiod-variation
Intraperiod-variation


$\mathbf{X}^{l}_{1D}$
Adaptive Aggregation
$\{\widehat{\mathbf{X}}^{l,1}_{1D}, \cdots, \widehat{\mathbf{X}}^{l,k}_{1D}\}$
1D Space
$\{\widehat{\mathbf{X}}^{l,1}_{2D}, \cdots, \widehat{\mathbf{X}}^{l,k}_{2D}\}$
Inception Block
2D Space
$\{\mathbf{X}^{l,1}_{2D}, \cdots, \mathbf{X}^{l,k}_{2D}\}$
1D Space
Time
$\mathbf{X}^{l-1}_{1D}$

Some Thinking

Building a pipeline.

Analysis in frequency is quite important

Which structure would be more suitable for casual in Time Series.