TILDE-Q: A TRANSFORMATION INVARIANT LOSS FUNCTION FOR TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series forecasting has gained increasing attention in the field of artificial intelligence due to its potential to address real-world problems across various domains, including energy, weather, traffic, and economy. While time-series forecasting is a well-researched field, predicting complex temporal patterns such as sudden changes in sequential data still poses a challenge with current models. This difficulty stems from minimizing L_p norm distances as loss functions, such as mean absolute error (MAE) or mean square error (MSE), which are susceptible to both intricate temporal dynamics modeling and signal shape capturing. Furthermore, these functions often cause models to behave aberrantly and generate uncorrelated results with the original time-series. Consequently, the development of a shape-aware loss function that goes beyond mere point-wise comparison is essential. In this paper, we examine the definition of shape and distortions, which are crucial for shape-awareness in time-series forecasting, and provide a design rationale for the shape-aware loss function. Based on our design rationale, we propose a novel, compact loss function called TILDE-Q (Transformation Invariant Loss function with Distance EQuilibrium) that considers not only amplitude and phase distortions but also allows models to capture the shape of time-series sequences. Furthermore, TILDE-Q supports the simultaneous modeling of periodic and nonperiodic temporal dynamics. We evaluate the efficacy of TILDE-Q by conducting extensive experiments under both periodic and nonperiodic conditions with various models ranging from naive to state-of-the-art. The experimental results show that the models trained with TILDE-Q surpass those trained with other metrics, such as MSE and DILATE, in various real-world applications, including electricity, traffic, illness, economics, weather, and electricity transformer temperature (ETT).

1 INTRODUCTION

Time-series forecasting has been a core problem across various domains, including traffic domain (Li et al., 2018; Lee et al., 2020), economy (Zhu & Shasha, 2002), and disease propagation analysis (Matsubara et al., 2014). One of the key challenges in time-series forecasting is the modeling of complex temporal dynamics (e.g., non-stationary signal and periodicity). Temporal dynamics, intuitively, shape, is the most emphasized keywords in time-series domains, such as rush hour of traffic data or abnormal usage of electricity (Keogh et al., 2003; Bakshi & Stephanopoulos, 1994; Weigend & Gershenfeld, 1994; Wu et al., 2021; Zhou et al., 2022).

Although deep learning methods are an appealing solution to model complex non-linear temporal dependencies and nonstationary signals, recent studies have revealed that even deep learning is often inadequate to model temporal dynamics. To properly model temporal dynamics, novel deep learning approaches, such as Autoformer (Wu et al., 2021) and FEDFormer (Zhou et al., 2022), have proposed input sequence decomposition. Still, they are trained with L_p norm-based loss function, which could not properly model the temporal dynamics, as shown in Fig. 1, (top). On the other hand, Le Guen & Thome (2019) attempt to model sudden changes in a timely and accurate manner with dynamic time warping (DTW), and Bica et al. (2020) adopt domain adversarial training to learn balanced representations, which is a treatment invariant representations over time. Le Guen & Thome (2019); Bica et al. (2020) try to capture the shape but still have some limitations, as depicted in Fig. 1 (middle), implying the need for further investigation of the shape.

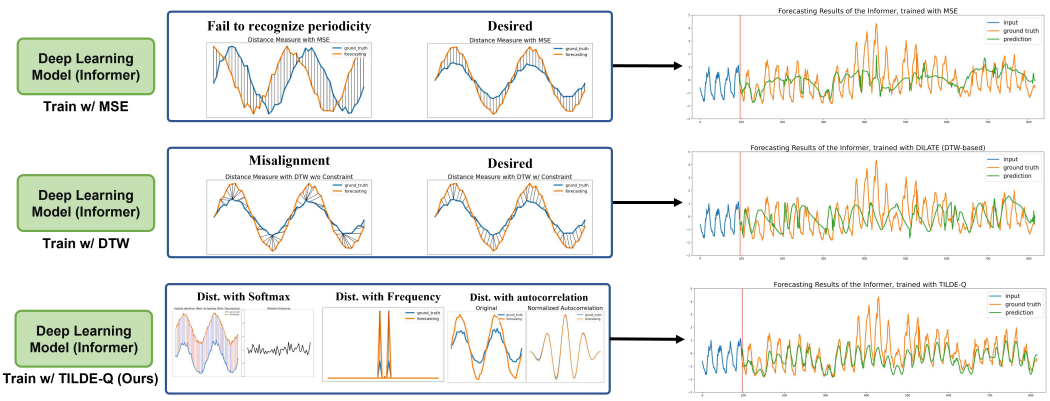

Figure 1: Ground-truth and forecasting results of Informer model with three training metrics, as shown in the blue box: (top) MSE, (middle) DTW-based, and (bottom) TILDE-Q loss function. (top, middle) The blue boxes indicates the original intention of loss function (desired) and misbehaviors.

The identification of **shape**, denoting the pattern in time-series data within a given time interval, plays an important role in addressing aforementioned limitation in time-series forecasting problem. It can provide valuable information, such as rise, drop, trough, peak, and plateau. We refer to the prediction as *informative* when it can appropriately model the shape. In real-world applications, including economics, informative prediction is invaluable for decision-making. To achieve such informative forecasting, a model should account for shape instead of solely aiming to forecast accurate value for each time step. However, existing methods inadequately consider the shape (Wu et al., 2021; Zhou et al., 2022; Bica et al., 2020; Le Guen & Thome, 2019). Moreover, deep learning model tends to opt for an *easy learning path* (Karras et al., 2019), yielding inaccurate and uninformative forecasting results disregarding the characteristics of time-series data. Fig. 1 illustrates three real forecasting results obtained with Informer (Zhou et al., 2021) and different training metrics. When the mean squared error (MSE) is used as an objective, the model aims to reduce the gap between prediction and ground truth for each time-step. This "point-wise" distance-based optimization has less ability to model shape, resulting in generating uninformative predictions regardless of temporal dynamics (Fig. 1 (top)); the model rarely provides information about the time-series. In contrast, if both gap and shape of the prediction and ground truth are taken into account, the model can achieve high accuracy with proper temporal dynamics, as shown in Fig. 1 (bottom). Consequently, time-series forecasting requires a loss function that consider both point-wise distance (i.e., traditional goal) and shape.

In this work, we aim to design a novel objective function that guides models in improving forecasting performance by learning shapes in time-series data. To design a shape-aware loss function, we review existing literature (Esling & Agon, 2012; Bakshi & Stephanopoulos, 1994; Keogh, 2003) and explore the concepts of *shapes* and *distortions* that impede appropriate measurement of similarity between two time-series data in terms of shapes (Sec. 3.1, Sec. 3.2, and Sec. 3.3). Based on our investigation, we propose the necessary conditions for constructing an objective function for shape-aware time-series forecasting (Sec. 4.1). Subsequently, we present a novel loss function, TILDE-Q (Transformation Invariant Loss function with Distance EQualibrium), which enables shape-aware representation learning by utilizing three loss terms that are invariant to distortions (Sec. 4.2). For evaluation, we conduct extensive experiments with state-of-the-art deep learning models with TILDE-Q. The experimental results indicate that TILDE-Q is model-agnostic and outperforms MSE and DILATE in MSE and shape-related metrics.

Contributions In summary, our study makes the following contributions. (1) We delve into the concept of shape awareness and distortion invariances in the context of time-series forecasting. By thoroughly investigating these distortions, we enhance our understanding of their impact on time-series forecasting problems. (2) We propose and implement TILDE-Q, which has invariances to three distortions and achieves shape-awareness, empowering informative forecasting in a timely manner. (3) We empirically demonstrate that the proposed TILDE-Q allows models to have higher accuracy compared to the models trained with other existing metrics, such as MSE and DILATE.

2 Related Work

2.1 Time-Series Forecasting

Many time-series forecasting methods are available, ranging from traditional models, such as ARIMA model (Box et al., 2015) and hidden Markov model (Pesaran et al., 2004), to recent deep learning models. In this section, we briefly describe the recent deep learning models for time-series forecasting. Motivated by the huge success of recurrent neural networks (RNNs) (Clevert et al., 2016; Li et al., 2018; Yu et al., 2017), many novel deep learning architectures have been developed for improving forecasting performance. To effectively capture long-term dependency, which is a limitation of RNNs, Stoller et al. (2020) have proposed convolutional neural networks (CNNs). However, it is required to stack lots of the same CNNs to capture long-term dependency (Zhou et al., 2021). Attention-based models, including Transformer (Vaswani et al., 2017) and Informer (Zhou et al., 2021), have been another popular research direction in time-series forecasting. Although these models effectively capture temporal dependencies, they incur high computational costs and often struggle to obtain appropriate temporal information (Wu et al., 2021). To cope with the problem, Wu et al. (2021); Zhou et al. (2022) have adopted the input decomposition method, which helps models better encode appropriate information. Other state-of-the-art models adopt neural memory networks (Kaiser et al., 2017; Sukhbaatar et al., 2015; Madotto et al., 2018; Lee et al., 2022), which refer to historical data stored in the memory to generate meaningful representation.

2.2 Training Metrics

Conventionally, mean squared error (MSE), L_p norm and its variants are mainstream metrics used to optimize forecasting models. However, they are not optimal for training forecasting models (Esling & Agon, 2012) because the time-series is temporally continuous. Moreover, the L_p norm provides less information about temporal correlation among time-series data. To better model temporal dynamics in time-series data, researchers have used differentiable, approximated dynamic time warping (DTW) as an alternative metric of MSE (Cuturi & Blondel, 2017; Abid & Zou, 2018; Mensch & Blondel, 2018). However, using DTW as a loss function results in temporal localization of changes being ignored. Recently, Le Guen & Thome (2019) have suggested DILATE, a training metric to catch sudden changes of nonstationary signals in a timely manner with smooth approximation of DTW and penalized temporal distortion index (TDI). To guarantee DILATE's operation in a timely manner, penalized TDI issues a harsh penalty when predictions showed high temporal distortion. However, the TDI relies on the DTW path, and DTW often showed misalignment because of noise and scale sensitivity. Thus, DILATE often loses its advantage with complex data, showing disadvantages at the training. In this paper, we discuss distortions and transformation invariances and design a new loss function that enables models to learn shapes in the data and produce noise-robust forecasting results.

3 Preliminary

In this section, we investigate common distortions focusing on the goal of time-series forecasting (i.e., modeling temporal dynamics and accurate forecasting). To clarify the concepts of time-series forecasting and related terms, we first define the notations and terms used (Sec. 3.1). We then discuss common distortions in time-series from the transformation perspective that need to be considered for building a shape-aware loss function (Sec. 3.2) and describe how other loss functions (e.g., dynamic time warping (DTW) and temporal distortion index (TDI)) handle shapes during learning (Sec. 3.3). We will discuss the conditions for effective time-series forecasting in the next session (Sec. 4.1).

3.1 Notations and Definitions

Let X_t denote a data point at a time step t. We define a time-series forecasting problem as follows:

Definition 3.1. *Given T-length historical time-series $\mathbf{X} = [X_{t-T+1}, \ldots, X_t], X_i \in \mathbb{R}^F$ at time i and a corresponding T'-length future time-series $\mathbf{Y} = [Y_{t+1}, \ldots, Y_{t+T'}], Y_i \in \mathbb{R}^C$, time-series forecasting aims to learn the mapping function $f : \mathbb{R}^{T \times F} \rightarrow \mathbb{R}^{T' \times C}$.*

To distinguish between the label (i.e., ground truth) and prediction time-series data, we note the label data as $\mathbf{Y}$ and prediction data as $\hat{\mathbf{Y}}$. Next, we set up two goals for time-series forecasting, which

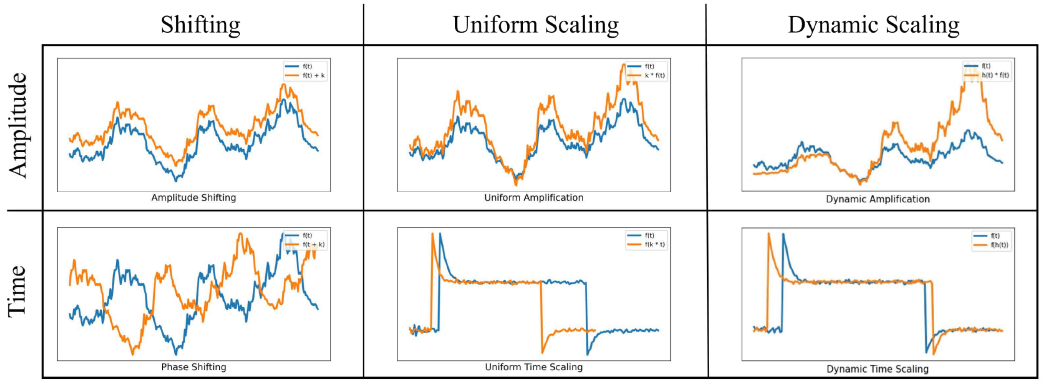

Figure 2: Example of the six distortions on the amplitude axis (top) and temporal axis (bottom).

require not only precise but also informative forecasting (Wu et al., 2021; Zhou et al., 2022; Le Guen & Thome, 2019) as follows:

- The mapping function f should be learnt to point-wisely reduce distance between $\hat{\mathbf{Y}}$ and $\mathbf{Y}$;
- The output $\hat{\mathbf{Y}}$ should have similar temporal dynamics with $\mathbf{Y}$.

Temporal dynamics are informative patterns in a time-series, such as rise, drop, peak, and plateau. The optimization for point-wise distance reduction is a conventional method used in the deep learning domain, which can be obtained using the MAE or MSE. However, in a real-world problem, such as traffic speed or stock market prediction, accurate forecasting of temporal dynamics is required. Esling & Agon (2012) also emphasized the measurement of temporal dynamics, as *"...allowing the recognition of perceptually similar objects even though they are not mathematically identical."* In this paper, we define temporal dynamics as follows:

Definition 3.2. *Temporal dynamics (or shapes) are informative periodic and nonperiodic patterns in time-series data.*

In this work, we aim to design a shape-aware loss function that satisfies both goals. To this end, we first discuss distortions that two time-series with similar shapes can have.

Definition 3.3. *Given two time-series* $\mathbf{F}$ *and* $\mathbf{G}$ *having similar shapes but not being mathematically identical, let* $\mathcal{H}$ *is transformation that satisfies* $\mathbf{F} = \mathcal{H}(\mathbf{G})$*. Then, the time-series* $\mathbf{F}$ *and* $\mathbf{G}$ *are considered to have a distortion, which can be represented by the transformation* $\mathcal{H}$*.*

A distortion can generally be classified as a temporal distortion (i.e., *warping*) or an amplitude distortion (i.e., *scaling*) depending on its dimension–time and amplitude. Existing distortions in the data lead to misbehavior of the model, as they distort the measurements to be inaccurate. For example, if we have two time-series $\mathbf{F}$ and $\mathbf{G} = \mathbf{F} + k$, which have similar shapes but different means, $\mathbf{G}$ could represent many temporal dynamics of $\mathbf{F}$. However, measurements often evaluate $\mathbf{F}$ and $\mathbf{G}$ as completely different signals and cause misguidance of the model in training (e.g., measuring the distance of $\mathbf{F}$ and $\mathbf{G}$ with MSE). As such, it is important to have measurements that consider a similar shape invariant to distortion. We define a measurement for distortion as:

Definition 3.4. *Let transformation* $\mathcal{H}$ *represent a distortion* H*. Then, we call measurement* $\mathcal{D}$ *invariant to* $\mathcal{H}$ *if* $\exists \delta > 0 : \mathcal{D}(\mathbf{T}, \mathcal{H}(\mathbf{T})) < \delta$ *for any time-series* $\mathbf{T}$*.*

3.2 Time-Series Distortions in Transformation Perspectives

Distortion, a gap between two similar time-series, affects shape capturing in time-series data. Thus, it is important to investigate different distortions and their impacts on representation learning aspects. There are six common time-series distortions that models encounter during learning (Esling & Agon, 2012; Batista et al., 2014; Berkhin, 2006; Warren Liao, 2005; Kerr et al., 2008)–Amplitude Shifting, Phase Shifting, Uniform Amplification, Uniform Time Scaling, Dynamic Amplification, and Dynamic Time Scaling. Next, we explain each common time-series distortion in terms of transformation with

an n-length time-series $\mathbf{F}(t) = [f(t_1), f(t_2), \ldots, f(t_n)]$, where t = $[t_1, t_2, \ldots, t_n]$. Fig. 2 presents example distortions, categorized by amplitude and time dimensions.

- *Amplitude Shifting* describes how much a time-series shifts against another time-series. This can be described with two time-series and the degree of shifting k: $\mathbf{G}(t) = \mathbf{F}(t) + k = [f(t_1) + k, \ldots, f(t_n) + k]$, where $k \in \mathbb{R}$ is constant.
- *Phase Shifting* is the same type of transformation (i.e., translation) as amplitude shifting, but it occurs along the temporal dimension. This distortion can be represented by two time-series functions with the degree of shift k: $\mathbf{G}(t) = \mathbf{F}(t + k) = [f(t_1 + k), \ldots, f(t_n + k)]$, where $k \in \mathbb{R}$ is constant. Cross-correlation (Paparrizos & Gravano, 2015; Vlachos et al., 2005) is the most popular measure method that is invariant to this distortion.
- *Uniform Amplification* is a transformation that changes the amplitude by multiplication of $k \in \mathbb{R}$. This distortion can be described with two functions and a multiplication factor k: $\mathbf{G}(t) = k \cdot \mathbf{F}(t) = [k \cdot f(t_1), \ldots, k \cdot f(t_n)]$.
- *Uniform Time Scaling* refers to a uniformly shortened or lengthened $\mathbf{F}(t)$ on the temporal axis. This distortion can be represented as $\mathbf{G}(t) = [g(t_1), \ldots, g(t_m)]$, where $g(t_i) = f(t_{\lceil k \cdot i \rceil})$ and $k \in \mathbb{R}^+$. Although Keogh et al. (2004) have proposed uniform time warping methods to handle this distortion, it still remains a challenging distortion type to measure because of the difficulty in identifying the scaling factor k without testing all possible cases (Keogh, 2003).
- *Dynamic Amplification* is any distortion that occurs through non-zero multiplication along the amplitude dimension. This distortion can be described as follows: $\mathbf{G}(t) = \mathbf{H}(t) \cdot \mathbf{F}(t) = [h(t_1) \cdot f(t_1), \ldots, h(t_n) \cdot f(t_n)]$ with function $h(t)$, such that $\forall_{t \in \mathbb{T}}, h(t) \neq 0$. Local amplification is representative of such distortions, which still remains challenging to solve.
- *Dynamic Time Scaling* refers to any transformation that dynamically lengthens or shortens signals along the temporal dimension, including local time scaling (Batista et al., 2014) and occlusion (Batista et al., 2014; Vlachos et al., 2003). It can be represented as follows: $\mathbf{G}(t) = \mathbf{F}(h(t)) = [f(h(t_1)), \ldots, f(h(t_n))]$, where $h(t)$ is a positive, strictly increasing function. DTW (Bellman & Kalaba, 1959; Berndt & Clifford, 1994; Keogh & Ratanamahatana, 2005) is the most popular technique invariant to this distortion. Das et al. (1997) have also introduced the longest common subsequence (LCSS) algorithm to tackle occlusion, noise, and outliers in this distortion.

Shape-aware clustering (Bellman & Kalaba, 1959; Batista et al., 2014; Paparrizos & Gravano, 2015; Berkhin, 2006; Warren Liao, 2005; Kerr et al., 2008) and classification (Xi et al., 2006; Batista et al., 2014; Srisai & Ratanamahatana, 2009) tasks that consider shapes have been extensively studied. However, only a few studies exist for time-series forecasting tasks, including Le Guen & Thome (2019) that utilize DTW and TDI for modeling temporal dynamics. Next, we describe the MSE and DILATE, proposed by Le Guen & Thome (2019), and discuss their invariance to distortions.

3.3 Distortion Handling in Current Time-Series Forecasting Objectives

Many measurement metrics have been used in the time-series forecasting domain, and those based on the L_p distance, including Euclidean distance, are widely used to handle time-series data. However, such metrics are not invariant to the aforementioned distortions (Ding et al., 2008; Le Guen & Thome, 2019) because of their point-wise mapping. In particular, since L_p distance compares the values per time step, it cannot handle temporal distortions appropriately and is vulnerable to data scaling. Le Guen & Thome (2019) have proposed a loss function called DILATE to overcome the inadequate characteristic in the L_p distance metric by recognizing temporal dynamics with DTW and TDI. In terms of transformation, DILATE handles dynamic time scaling, especially local time scaling with DTW, and phase shifting with penalized TDI, defined as follows:

$$\mathcal{L}_{dilate}(\hat{y}_i, y_i) := -\gamma \log \Big(\sum_{\mathbf{A} \in \mathcal{A}_{k,k}} e^{-\frac{\langle \mathbf{A}, \alpha \Delta(\hat{y}_i, y_i) + (1-\alpha)\mathbf{\Omega} \rangle}{\gamma}} \Big),$$

where $\mathbf{A}$, $\Delta(\hat{y}_i, y_i)$, $\mathbf{\Omega}$ are the warping path, cost matrix, and penalization matrix, respectively.

While DILATE performs better than existing methods, it has a limitation from the perspective of invariance. DILATE highly depends on DTW, which allows for the dynamic alignment of the

time-series for a predefined window. In such windows, DTW can align the signal regardless of its information (e.g., periodicity). As a result, the model creates misbehavior that can cheat DTW within the window, as shown in Fig. 1 middle. DTW's scale and noise sensitivity are also problematic. DTW computes the Euclidean distance of two time-series after its temporal alignment in dynamic programming, and the alignment relies on the distance function. Consequently, the dynamic alignment of DTW can be properly achieved only when the two time-series have the same range (Esling & Agon, 2012; Bellman & Kalaba, 1959). This means that it hardly achieves invariance to amplitude distortion without appropriate pre-processing. Gong & Chen (2017) also show that DTW poorly matches the prediction and target (i.e., ground truth) time-series with amplitude shifting. Even when the target time-series is aligned with normalization, the appropriate alignment of the prediction and target time-series cannot be guaranteed because of DTW's high sensitivity to noise. As a result, DILATE can generate poor alignment results, which can cause wrong TDI optimization, producing incorrect results and instability during the optimization steps. To design an effective shape-aware loss function, we must understand the measures and in which cases they have transformation invariances. In the next section, we interpret transformations from a time-series forecasting viewpoint and discuss the types of transformations that should be considered in objective function design.

4 METHODS

In this section, we discuss and propose the design rationale for the shape-aware loss function (Sec. 4.1). Based on the design rationale, we implement a novel loss function, TILDE-Q (a Transformation Invariant Loss function with Distance EQuilibrium), which allows models to perform shape-aware time-series forecasting based on three distortion invariances.

4.1 TRANSFORMATION INVARIANCES IN TIME-SERIES FORECASTING

In the time-series domain, data often have various distortions; thus, measurements need to satisfy numerous transformation invariances for meaningfully modeling temporal dynamics. As discussed in Sec. 3.1, we set the goals of time-series forecasting as (1) point-wisely reducing the gap between the prediction and target time-series and (2) preserving the temporal dynamics of the target time-series. To satisfy both of them, we have to consider (1) a method that does not negatively impact on the traditional goal of accurate time-series forecasting and (2) distortions that play a crucial role in capturing the temporal dynamics of the target time-series. In this section, we review all six distortions based on whether their corresponding invariance is feasible to be a loss function for time-series forecasting, discuss the loss function's benefits and trade-offs, and identify appropriate distortions to be considered in time-series forecasting.

Amplitude Shifting In a wide range of situations, it is beneficial to capture the trends of time-series sequences despite shifts in amplitude. Thus, being invariant to amplitude shifting in a loss function is highly advantageous in time-series forecasting: (1) shape awareness invariant to amplitude shifting, (2) accurate deviation of values in modeling, and (3) effective on-time prediction of the peak or sudden changes. To guarantee an amplitude shifting invariance in the optimization stage, the loss function should induce an equal gap k between the prediction and ground truth data in each step. Specifically, the loss function considering amplitude shifting should satisfy:

$$\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = 0 \Leftrightarrow \forall_{i \in [1,\dots,n]}, d(y_i, \hat{y}_i) = k, \tag{1}$$

where $k \in \mathbb{R}$ is an arbitrary and equal gap, and $d(y_i, \hat{y}_i)$ is a signed distance with a boundary $y_i > \hat{y}_i$. By allowing tolerance between the prediction and target time-series, models can follow trends in time-series instead of predicting exact values point-wisely. In short, unlike existing loss functions, which handle only point-wise distance (e.g., DTW), we should deal with both point-wise distance and its relational distance values to guarantee amplitude shifting.

Phase Shifting There are some forecasting tasks whose main objectives concern accurate forecasting of peaks and periodicity in time-series (e.g., heartbeat data and stock price data). For such tasks, phase shifting invariance is an optimal solution for (1) modeling periodicity, regardless of the translation on the temporal axis, and (2) having precise statistics with shapes, such as peak and plateau values. To be invariant to phase shifting, the loss function should satisfy

$$\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = 0 \Leftrightarrow \mathbf{Y}, \hat{\mathbf{Y}} \text{ have the same dominant frequency.} \tag{2}$$

Note that Eq. 2 allows a similar shape as the target time-series in forecasting, not exactly the same shape (e.g., $\sin(x)$ and $2\sin(x + x_0)$ with the same dominant frequency).

Uniform Amplification This proposition can be utilized in the case of sparse data that contains a significant number of zeros. By adopting uniform amplification invariance, models are able to focus on non-zero sequences, whereas this proposition allows models to receive less penalty in zero sequences. Since it guarantees shape awareness with a multiplication factor in a timely manner, as shown in Fig. 2, invariance for uniform amplification fits well. To have a model trained with uniform amplification invariance, the loss function should satisfy the following proposition:

$$\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = 0 \Leftrightarrow \forall_{i \in [1,\dots,n]}, \frac{y_i}{\hat{y}_i} = k(\hat{y}_i \neq 0). \tag{3}$$

Uniform Time Scaling, Dynamic Amplification, and Dynamic Time Scaling After careful consideration, we conclude that uniform time scaling, dynamic amplification, and dynamic time scaling are incompatible for optimization. the reasons are described below.

To achieve invariance for uniform time scaling, the loss function should satisfy below:

$$\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = 0 \Leftrightarrow \exists c \in \mathbb{Z}^+, \text{where } \{c | y_i = \hat{y}_{ci}\} \cup \{c | y_{ci} = \hat{y}_i\} \forall i \in [0, 1, \dots, T']. \tag{4}$$

This proposition will negatively influence the original temporal dynamics, considering that it creates the tolerance for mispredicting periodicity (e.g., daily periodic signals) and cannot identify events (e.g., abrupt changing values) in a timely manner. In summary, it hinders models from capturing shapes and corrupts periodic information.

For both dynamic amplification and time scaling, the loss functions are zero for all pairs when there is no limit for tolerance. Formally, the proposition for dynamic amplification invariance is as follows:

$$\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = 0 \Leftrightarrow \forall c_i \in \mathbb{R} : y_i = c_i \hat{y}_i,$$

If a loss function satisfies this proposition without bound for c_i, it is always zero because there always exists $c_i = y_i / \hat{y}_i$, except $\hat{y}_i = 0$. Therefore, it is not able to provide any information because all random values could be an optimal solution. The same situation happens for the dynamic time scaling if we do not limit the window. Consequently, all three objectives–uniform time scaling, dynamic amplification, and dynamic time scaling are unsuitable to be objectives in time-series forecasting.

4.2 TILDE-Q: Transformation Invariant Loss function with Distance EQuilibrium

To build a transformation invariant loss function, we need to design a loss function that satisfies the proposition for amplitude shifting (Eq. 1), phase shifting (Eq. 2), and uniform amplification shifting invariance (Eq. 3), as discussed in Sec. 4.1. Furthermore, the loss function should guarantee a small L_p norm between prediction and label, which is the traditional goal of forecasting. Both conditions are hard to simultaneously satisfy by existing loss functions, such as the MSE or DILATE. To handle all three distortions while considering traditional goal, we build three objective functions (*a.shift*, *phase*, and *amp* losses) that can achieve one or more invariance by using softmax, Fourier coefficient, and autocorrelation to design a loss function.

Amplitude Shifting Invariance with Softmax (Amplitude Shifting) To strengthen amplitude shifting invariance, we design a loss function that satisfies Eq. 1. This means that $d(y_i, \hat{y}_i)$ must have the same value for all i. To satisfy this condition, we utilize the softmax function:

$$\mathcal{L}_{a.shift}(\mathbf{Y}, \hat{\mathbf{Y}}) = T' \sum_{i=1}^{T'} |\frac{1}{T'} - \text{Softmax}(d(y_i, \hat{y}_i))|, \tag{5}$$

where T', Softmax, and $d(\cdot, \cdot)$ are the sequence length, softmax function, and signed distance function, respectively. Because softmax produces the proportion of each value, it can obtain the optimal solution only when it satisfies Eq. 1. Since Softmax outputs the relative values, it could handle any gap k.

Table 1: Experimental results on six real-world datasets in multivariate time-series forecasting setting with prediction lengths $T' = \{24, 36, 48, 60\}$ for ILI and $T' = \{96, 192, 336, 720\}$ for others. The results are averaged from all prediction lengths. We set input sequence length $T = 96$ except ILI dataset. For ILI dataset, we set input sequence length $T = 36$. *Improved* means the average improvements of TILDE-Q over the model trained with MSE. We have colored the best training metric in **red**.

Model	iTransformer				PatchTST				Crossformer				TimesNet				DLinear				FEDformer				NSformer				Autoformer			
Methods	MSE		TILDE-Q		MSE		TILDE-Q		MSE		TILDE-Q		MSE		TILDE-Q		MSE		TILDE-Q		MSE		TILDE-Q		MSE		TILDE-Q		MSE		TILDE-Q	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETT*	0.408	0.412	**0.403**	**0.405**	**0.387**	0.400	0.387	**0.397**	0.535	0.521	**0.427**	**0.439**	0.415	0.419	**0.401**	**0.411**	0.404	0.408	**0.401**	**0.400**	0.461	0.459	**0.438**	**0.451**	0.533	0.470	**0.512**	**0.465**	0.586	0.516	**0.541**	**0.492**
Electricity	0.179	0.269	**0.175**	**0.266**	0.204	0.291	**0.203**	**0.282**	0.188	0.284	**0.181**	**0.278**	0.195	0.295	**0.192**	**0.292**	0.225	0.319	**0.212**	**0.294**	0.227	0.337	**0.224**	**0.333**	0.196	0.295	**0.194**	**0.293**	0.250	0.351	**0.232**	**0.338**
Traffic	0.428	0.282	**0.426**	**0.281**	0.555	0.362	**0.514**	**0.335**	0.550	0.304	**0.540**	**0.296**	0.620	0.336	**0.600**	**0.328**	0.672	0.418	**0.667**	**0.399**	0.610	0.378	**0.606**	**0.376**	0.644	0.355	**0.637**	**0.351**	0.637	0.395	**0.619**	**0.386**
Weather	0.260	0.281	**0.257**	**0.274**	0.262	0.281	**0.258**	**0.280**	0.259	0.315	**0.248**	**0.301**	0.259	0.287	**0.256**	**0.282**	0.268	0.317	**0.266**	**0.306**	0.309	0.360	**0.302**	**0.342**	0.312	0.323	**0.310**	**0.317**	0.366	0.396	**0.346**	**0.376**
Exchange	0.389	0.421	**0.379**	**0.415**	0.365	0.410	**0.364**	**0.403**	0.943	0.711	**0.833**	**0.660**	**0.403**	**0.436**	0.407	0.438	0.323	0.392	**0.301**	**0.379**	0.554	0.515	**0.524**	**0.499**	0.546	0.488	**0.474**	**0.457**	0.532	0.518	**0.494**	**0.493**
ILI	2.333	0.984	**2.220**	**0.949**	2.253	0.933	**2.143**	**0.903**	3.724	1.281	**3.297**	**1.202**	2.346	0.963	**2.083**	**0.899**	2.815	1.150	**2.475**	**1.071**	3.307	1.276	**3.113**	**1.225**	2.613	1.024	**2.032**	**0.921**	3.327	1.261	**3.199**	**1.241**
Improved	-	-	2.08%	1.77%	-	-	2.43%	2.76%	-	-	8.85%	6.38%	-	-	3.25%	2.21%	-	-	4.48%	4.67%	-	-	3.42%	2.59%	-	-	7.02%	3.5%	-	-	5.69%	3.68%

Invariances with Fourier Coefficients (Phase Shifting) As discussed in Sec. 4.1, a potential method that can be used to obtain phase shifting invariance is the use of Fourier coefficients. According to the literature (NG & GOLDBERGER, 2007), the original time-series can be reconstructed with a few dominant frequencies. Thus, we utilize the gap between dominant Fourier coefficients of ground truth and prediction as our objective function for achieving phase shifting invariance. For the other frequencies, we use the norm of the prediction sequence to reduce the value of the Fourier coefficient. Consequently, this loss function keeps the temporal dynamics of the original time series (i.e., dominant frequencies) and enables noise robustness by reducing white noises in non-dominant frequencies. We achieve phase shifting invariance by optimizing the following loss function:

$$\mathcal{L}_{phase}(\mathbf{Y}, \hat{\mathbf{Y}}) = \begin{cases} ||\mathcal{F}(\mathbf{Y}) - \mathcal{F}(\hat{\mathbf{Y}})||_p, & \text{dominant freq.} \\ ||\mathcal{F}(\hat{\mathbf{Y}})||_p, & \text{otherwise} \end{cases} \tag{6}$$

where $||\cdot||_p$ is the L_p norm. To obtain the dominant frequency terms, we calculate the norm of the Fourier coefficient for each frequency and filter them with the squared root of sequence length, $\sqrt{T'}$. We also guarantee the minimum number of dominant frequencies as $\sqrt{T'}$. This loss function obtains uniform amplification invariance through the application of a normalization technique to Fourier coefficients. For example, $\sin x$ and $c \cdot \sin x$ have the same Fourier coefficients if appropriately normalized. In summary, from Eq. 6, we can obtain (1) invariance for phase shifting, (2) invariance for uniform amplification, and (3) robustness to noise.

Invariances with Autocorrelation (Uniform Amplification) Although Fourier coefficients can be considered a reasonable solution to determine the periodicity of the target time-series, they are not completely invariant to phase shifting for three reasons: (1) the data statistics (e.g., mean and variance) keep changing, (2) such changing statistics also cause changes in Fourier coefficients even at the same frequency, and (3) objectives only with a norm of Fourier coefficient cannot fully represent the original time-series. Thus, we introduce an objective based on normalized cross-correlation, which satisfies Eq. 2 for a periodic signal:

$$\mathcal{L}_{amp}(\mathbf{Y}, \hat{\mathbf{Y}}) = ||R(\mathbf{Y}, \mathbf{Y}) - R(\mathbf{Y}, \hat{\mathbf{Y}})||_p, \tag{7}$$

where $R(\cdot, \cdot)$ is a normalized cross-correlation function. This loss function helps predicted sequences mimic label sequences by calculating the difference between the autocorrelation of the label sequences and the cross-correlation between the label and predicted sequences. Therefore, the label and prediction have similar temporal dynamics, regardless of phase shifting or uniform amplification.

In summary, we introduce TILDE-Q, combining Eq. 5, Eq. 6, and Eq. 7 as follows:

$$\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = \alpha \mathcal{L}_{a.shift}(\mathbf{Y}, \hat{\mathbf{Y}}) + (1 - \alpha)\mathcal{L}_{phase}(\mathbf{Y}, \hat{\mathbf{Y}}) + \gamma \mathcal{L}_{amp}(\mathbf{Y}, \hat{\mathbf{Y}}), \tag{8}$$

where $\alpha \in [0, 1]$ and γ are hyperparameters.

5 EXPERIMENTS

In this section, we present the results of our comprehensive experiments, demonstrating the effectiveness of TILDE-Q and the importance of transformation invariance.

Table 2: Experimental results of short-term time-series forecasting on the three datasets with sequence-to-sequence GRU model. We have colored the best training metric in **red** and the second best underlined.

Methods	GRU + MSE				GRU + DILATE				GRU + TILDE-Q			
Eval	MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS
Synthetic	**0.0107**	3.5080	**1.0392**	0.3523	0.0130	3.4005	1.1242	**0.3825**	0.0119	**3.2873**	1.1564	0.3811
ECG5000	0.2152	1.9718	0.8442	0.7743	0.8270	3.9579	2.0281	0.4356	**0.2141**	**1.9575**	**0.7714**	**0.7773**
Traffic	**0.0070**	1.4628	0.2343	0.7209	0.0095	1.6929	0.2814	0.6806	0.0072	**1.4600**	**0.2276**	**0.7220**

Experimental Setup We conduct the experiments with eight state-of-the-art models–Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), NSformer (Liu et al., 2022), DLinear (Zeng et al., 2023), TimesNet (Wu et al., 2023), Crossformer (Zhang & Yan, 2023), PatchTST Nie et al. (2023), and iTransformer (Liu et al., 2023) and one basic GRU model. For model training, we use seven real-world datasets–ECL, ETT, Electricity, Traffic, Weather, Exchange, and Weather–and one synthetic dataset, Synthetic. We repeat each experiment with a model and dataset 10 times in combination with two different objective functions. Appendix B provides detailed explanations of the datasets, hyperparameter settings, model, and source code. We also provide additional qualitative results in Appendix.

Evaluation Metrics In this experiment, we evaluate TILDE-Q with three evaluation metrics: mean absolute error (MAE), mean squared error (MSE), dynamic time warping (DTW), and its corresponding temporal distortion index (TDI), all of which are referred from Le Guen & Thome (2019). As DTW is sensitive to noise and generates incorrect paths when one of the time-series data is noisy (as discussed in Sec. 3.3), we additionally use the longest common subsequence (LCSS) for comparison, which is more robust to outliers and noise (Esling & Agon, 2012). The longer the matched subsequences, the higher the LCSS score will be achieved in modeling the shapes. For state-of-the-art models, we report the MAE and MSE. For detailed results, including forecasting results for different prediction lengths, please refer to Appendix C.

Experimental Results and Analysis Table 2 shows the results of the short-term forecasting performance of the GRU model optimized with the MSE, DILATE, and TILDE-Q metrics. With the Synthetic dataset, each metric used shows its own benefits. This result indicates that loss functions with shape similarity or MSE have their specialty for shape and exact value, respectively. It also means a better MSE does not guarantee a better solution for temporal dynamics. Moreover, since the model is evaluated with real-world datasets, it is revealed that TILDE-Q outperforms other objective functions in most evaluation metrics. These results indicate that our approach to learning shapes in time-series data achieves better results than existing methods for forecasting. DILATE does not show impressive performance with ECG5000 due to its high sensitivity to noise, as discussed in Sec. 3.3. Table 1 summarizes the comprehensive experimental results obtained with the eight state-of-the-art models. Compared to MSE baseline, TILDE-Q makes particularly better prediction among all the models, even for the DLinear (Zeng et al., 2023), which consists of two one-layer linear layers. For the Autoformer and NSformer, TILDE-Q makes significant improvement around 5%, making the models recognize additional shape-related information beyond the frequency-based terms. For the recent models (i.e., iTransformer and PatchTST) that interpret input signals with embedding or patching, TILDE-Q is less beneficial than the other models. Crossformer makes the most impressive results with 8.85% performance improvement. This improvement is caused by Crossformer's design, which particularly focuses on resolving inter-domain dependency among multivariate time series. TILDE-Q is able for Crossformer to recognize the cross-time dependency (i.e., shape and temporal changes) better, indicating the importance of both temporal and inter-domain behaviors. This insight reveals possible future research investigating loss function related to causality and inter-domain dependency.

6 Conclusion and Future Work

We propose TILDE-Q that allows shape-aware time-series forecasting in a timely manner. To design TILDE-Q, we review existing transformations in time-series data and discuss the conditions that ensure transformation invariance during optimization tasks. The designed TILDE-Q is invariant to amplitude shifting, phase shifting, and uniform amplification, ensuring a model better captures shapes in time-series data. To prove the effectiveness of TILDE-Q, we conduct comprehensive experiments with state-of-the-art models and real-world datasets. The results indicate that the model trained with TILDE-Q generates more timely, robust, accurate, and shape-aware forecasting in both short-term and long-term forecasting tasks. We conjecture that this work can facilitate future research on transformation invariances and shape-aware forecasting.

7 Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abid, A. and Zou, J. Y. Learning a warping distance from unlabeled time series using sequence autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10568–10578, 2018.

Bakshi, B. and Stephanopoulos, G. Representation of process trends—iv. induction of real-time patterns from operating data for diagnosis and supervisory control. *Computers & Chemical Engineering*, 18(4):303–332, 1994.

Batista, G. E. A. P. A., Keogh, E. J., Tataw, O. M., and de Souza, V. M. A. CID: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, 28(3): 634–669, 2014. doi: 10.1007/s10618-013-0312-3.

Bellman, R. and Kalaba, R. On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2):1–9, 1959.

Berkhin, P. A survey of clustering data mining techniques. In *Grouping Multidimensional Data - Recent Advances in Clustering*, pp. 25–71. Springer, 2006.

Berndt, D. J. and Clifford, J. Using dynamic time warping to find patterns in time series. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, pp. 359–370. AAAI Press, 1994.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time series analysis: forecasting and control*. John Wiley, 2015.

Clevert, D., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). In *Proceedings of the International Conference on Learning Representations*, 2016.

Cuturi, M. and Blondel, M. Soft-dtw: A differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, ICML'17, pp. 894–903, 2017.

Das, G., Gunopulos, D., and Mannila, H. Finding similar time series. In *Principles of Data Mining and Knowledge Discovery*, pp. 88–100, 1997.

Dau, H. A., Bagnall, A., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., and Keogh, E. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6): 1293–1305, 2019.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

Esling, P. and Agon, C. Time-series data mining. *ACM Computing Surveys*, 45(1), 2012.

Gong, Z. and Chen, H. Dynamic state warping. *CoRR*, abs/1703.01141, 2017.

Kaiser, L., Nachum, O., Roy, A., and Bengio, S. Learning to remember rare events. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Keogh, E. J. Efficiently finding arbitrarily scaled patterns in massive time series databases. In *Knowledge Discovery in Databases: PKDD 2003*, volume 2838 of *Lecture Notes in Computer Science*, pp. 253–265, 2003.

Keogh, E. J. and Ratanamahatana, C. A. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, 2005.

Keogh, E. J., Lin, J., and Truppel, W. Clustering of time series subsequences is meaningless: Implications for previous and future research. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 115–122. IEEE Computer Society, 2003.

Keogh, E. J., Palpanas, T., Zordan, V. B., Gunopulos, D., and Cardle, M. Indexing large human-motion databases. In *Proceedings of the International Conference on Very Large Data Bases*, pp. 780–791, 2004.

Kerr, G., Ruskin, H., Crane, M., and Doolan, P. Techniques for clustering gene expression data. *Computers in Biology and Medicine*, 38(3):283–293, 2008.

Le Guen, V. and Thome, N. Shape and time distortion loss for training deep time series forecasting models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Lee, C., Kim, Y., Jin, S., Kim, D., Maciejewski, R., Ebert, D., and Ko, S. A visual analytics system for exploring, monitoring, and forecasting road traffic congestion. *IEEE Transactions on Visualization and Computer Graphics*, 26(11):3133–3146, 2020. doi: 10.1109/TVCG.2019.2922597.

Lee, H., Jin, S., Chu, H., Lim, H., and Ko, S. Learning to remember patterns: Pattern matching memory networks for traffic forecasting. In *International Conference on Learning Representations*, 2022.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2018.

Liu, Y., Wu, H., Wang, J., and Long, M. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, 2022.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Madotto, A., Wu, C., and Fung, P. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 1468–1478, 2018.

Matsubara, Y., Sakurai, Y., van Panhuis, W. G., and Faloutsos, C. FUNNEL: automatic mining of spatially coevolving epidemics. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 105–114. ACM, 2014.

Mensch, A. and Blondel, M. Differentiable dynamic programming for structured prediction and attention. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3462–3471, 2018.

NG, J. and GOLDBERGER, J. J. Understanding and interpreting dominant frequency analysis of af electrograms. *Journal of Cardiovascular Electrophysiology*, 18(6):680–685, 2007.

Nie, Y., H. Nguyen, N., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Paparrizos, J. and Gravano, L. K-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pp. 1855–1870, 2015. doi: 10.1145/2723372.2737793.

Pesaran, M., Pettenuzzo, D., and Timmermann, A. Forecasting time series subject to multiple structural breaks. Cambridge Working Papers in Economics 0433, Faculty of Economics, University of Cambridge, 2004.

Srisai, D. and Ratanamahatana, C. A. Efficient time series classification under template matching using time warping alignment. In *Proceedings of the International Conference on Computer Sciences and Convergence Information Technology*, pp. 685–690, 2009.

Stoller, D., Tian, M., Ewert, S., and Dixon, S. Seq-u-net: A one-dimensional causal u-net for efficient sequence modelling. In Bessiere, C. (ed.), *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2893–2900. ijcai.org, 2020.

Sukhbaatar, S., szlam, a., Weston, J., and Fergus, R. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Vlachos, M., Hadjieleftheriou, M., Gunopulos, D., and Keogh, E. Indexing multi-dimensional time-series with support for multiple distance measures. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pp. 216–225, 2003.

Vlachos, M., Yu, P. S., and Castelli, V. On periodicity detection and structural periodic similarity. In *Proceedings of the SIAM International Conference on Data Mining*, pp. 449–460, 2005.

Warren Liao, T. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

Weigend, A. S. and Gershenfeld, N. A. *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley, 1994. ISBN 0-201-62601-2.

Wu, H., Xu, J., Wang, J., and Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, 2021.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Xi, X., Keogh, E., Shelton, C., Wei, L., and Ratanamahatana, C. A. Fast time series classification using numerosity reduction. In *Proceedings of the International Conference on Machine Learning*, ICML '06, pp. 1033–1040. Association for Computing Machinery, 2006.

Yu, F., Koltun, V., and Funkhouser, T. A. Dilated residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 636–644. IEEE Computer Society, 2017.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Zhang, Y. and Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., and Jin, R. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286, 2022.

Zhu, Y. and Shasha, D. Statstream: Statistical monitoring of thousands of data streams in real time. In *Proceedings of the International Conference on Very Large Databases*, pp. 358–369. Morgan Kaufmann, 2002.