

INFORMATION-AWARE TIME SERIES META-CONTRASTIVE LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

Various contrastive learning approaches have been proposed in recent years and achieve significant empirical success. While effective and prevalent, contrastive learning has been less explored for time series data. A key component of contrastive learning is to select appropriate augmentations imposing some priors to construct feasible positive samples, such that an encoder can be trained to learn robust and discriminative representations. Unlike image and language domains where “desired” augmented samples can be generated with the rule of thumb guided by prefabricated human priors, the ad-hoc manual selection of time series augmentations is hindered by their diverse and human-unrecognizable temporal structures. How to find the desired augmentations of time series data that are meaningful for given contrastive learning tasks and datasets remains an open question. In this work, we address the problem by encouraging both high *fidelity* and *variety* based upon information theory. A theoretical analysis leads to the criteria for selecting feasible data augmentations. On top of that, we employ the meta-learning mechanism and propose an information-aware approach, InfoTS, that adaptively selects optimal time series augmentations for contrastive representation learning. The meta-learner and the encoder are jointly optimized in an end-to-end manner to avoid sub-optimal solutions. Experiments on various datasets show highly competitive performance with up to 11.4% reduction in MSE on forecasting task and up to 2.8% relative improvement in accuracy on classification task over the leading baselines.

1 INTRODUCTION

Time series data in the real world is high dimensional, unstructured, and complex with unique properties, leading to challenges for data modeling (Yang & Wu, 2006). In addition, without human recognizable patterns, it is much harder to label time series data than images and languages in real-world applications. These labeling limitations hinder deep learning methods, which typically require a huge amount of labeled data for training, been applied on time series data (Eldele et al., 2021). Representation learning learns a fixed-dimension embedding from the original time series that keeps their inherent features. Comparing to the raw time series data, these representations are with better transferability and generalization capacity. To deal with labeling limitations, contrastive learning methods have been widely adopted in various domains for their soaring performance on representation learning, including vision, language, and graph-structured data (Chen et al., 2020; Xie et al., 2019; You et al., 2020). In a nutshell, contrastive learning methods typically train an encoder to map instances to an embedding space where dissimilar (negative) instances are easily distinguishable from similar (positive) ones and model predictions to be invariant to small noise applied to either input examples or hidden states.

Despite being effective and prevalent, contrastive learning has been less explored in the time series domain (Eldele et al., 2021; Franceschi et al., 2019; Fan et al., 2020; Tonekaboni et al., 2021). Existing contrastive learning approaches often involve a specific data augmentation strategy that creates novel and realistic-looking training data without changing its label to construct positive alternatives for any input sample. Their success relies on carefully designed rules of thumb guided by domain expertise. Routinely used data augmentations for contrastive learning are mainly designed for image and language data, such as color distortion, flip, word replacement, and back-translation (Chen et al., 2020; Luo et al., 2021). These augmentation techniques generally do not apply to time series data.

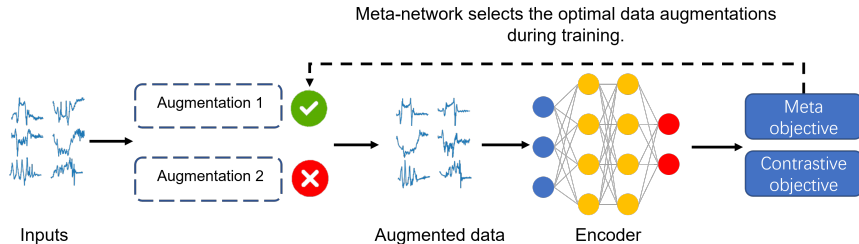


Figure 1: InfoTS is composed of three parts: (1) candidate transformation that generates different augmentations of the original inputs, (2) a meta-network that selects the optimal augmentations, (3) an encoder that learns representations of time series instances. The meta-network is learned in tandem with contrastive encoder learning.

Recently, some researchers propose augmentations for time series to enhance the size and quality of the training data (Wen et al., 2021). For example, Eldele et al. (2021) and Fan et al. (2020) propose to adopt jittering, scaling, and permutation strategies to generate augmented instances. Franceschi et al. (2019) extracts subsequences for data augmentation. In spite of the current progress, existing methods have two main limitations. First, unlike images with human recognizable features, time series data are often associated with inexplicable underlying patterns. Strong augmentation such as permutation may ruin such patterns and consequently, the model will mistake the negative handcrafts for positive ones. While weak augmentation methods such as jittering may generate augmented instances that are too similar to the raw inputs to be informative enough for contrastive learning. On the other hand, time series datasets from different domains may have diverse nature. Adapting a universal data augmentation method, such as subsequence in Xie et al. (2019), for all datasets and tasks leads to sub-optimal performances. Other works follow empirical rules to select suitable augmentations from expensive trial-and-error. Akin to hand-crafting features, hand-picking choices of data augmentations are undesirable from the learning perspective. The diversity and heterogeneity of real-life time series data further hinder these methods away from wide applicability.

To address the challenges, we first introduce the criteria for selecting good data augmentations in contrastive learning. Data augmentation benefits generalizable, transferable, and robust representation learning by correctly extrapolating the input training space to a larger region (Wilk et al., 2018). The positive instances enclose a discriminative zone in which all the data points should be similar to the original instance. The desired data augmentations for contrastive representation learning should have both *high fidelity* and *high variety*. High fidelity encourages the augmented data to maintain the semantic identity that is invariant to transformations (Wilk et al., 2018). For example, if the downstream task is classification, then the generated augmentations of inputs should be class-preserving. Meanwhile, generating augmented samples with high variety benefits representation learning by increasing the generalization capacity (Chen et al., 2020). From the motivation, we theoretically analyze the information flows in data augmentations based upon information theory and derive the criteria for selecting desired time series augmentations. Due to the inexplicability in practical time series data, we assume that the semantic identity is presented by the target in the downstream task. Thus, high fidelity can be achieved by maximizing the mutual information between the downstream label and the augmented data. A one-hot pseudo label is assigned to each instance in the unsupervised setting when downstream labels are unavailable. These pseudo labels encourage augmentations of different instances to be distinguishable from each other. We show that data augmentations preserving these pseudo labels can add new information without decreasing the fidelity. Concurrently, we maximize the entropy of augmented data conditional on the original instances to increase the variety of data augmentations.

Based on the derived criteria, we propose an adaptive data augmentation method, InfoTS (as shown in Figure 1), by employing a meta-learning mechanism to avoid ad-hoc choices or painstakingly trial-and-error tuning. Specifically, we utilize a meta-network to learn the augmentation prior in tandem with contrastive learning. The meta-learner automatically selects optimal augmentations from candidate augmentations to generate feasible positive samples. Along with random sampled negative instances, augmented instances are then fed into a time series encoder to learn representations in a contrastive manner. With a reparameterization trick, the meta-network can be efficiently optimized with back-propagation based upon the proposed criteria. Therefore, the meta-network can automatically select data augmentations in a per dataset and per learning task manner without resorting to expert knowledge or tedious downstream validation. Our main contributions include:

- We propose criteria to guide the selection of data augmentations for contrastive time series representation learning without prefabricated knowledge.
- We propose a meta-learning based method to automatically select feasible data augmentations for different time series datasets, which can be efficiently optimized with back-propagation.
- We empirically verify the effectiveness of the proposed criteria to find optimal data augmentations. Extensive experiments demonstrate that InfoTS can achieve highly competitive performance with up to 11.4% reduction in MSE on forecasting task and up to 2.8% relative improvement in accuracy on classification task over the leading baselines.

2 METHODOLOGY

2.1 NOTATIONS AND PROBLEM DEFINITION

A time series instance x has dimension $T \times F$, where T is the length of sequence and F is the dimension of features. Given a set of time series instances \mathbb{X} , we aim to learn an encoder $f_{\theta}(x)$ that maps each instance x to a fixed-length vector $\mathbf{z} \in \mathbb{R}^D$, where θ is the learnable parameters of the encoder network and D is the dimension of representation vectors. In semi-supervised settings, each instance x in the labelled set $\mathbb{X}_L \subseteq \mathbb{X}$ is associated with a label y for the downstream task. Specially, $\mathbb{X}_L = \mathbb{X}$ holds in the fully supervised setting. In the work, we use the Sans-serif style lowercase letters, such as x , to denote random time series variables and italic lowercase letters, such as x , for sampled instances.

2.2 WHAT ARE GOOD AUGMENTATIONS FOR CONTRASTIVE LEARNING?

The goal of data augmentation for contrastive learning is to create realistically rational instances that maintain semantics through different transformation approaches. Unlike instances in vision and language domains, the underlying semantics of time series data is not recognizable to human, making it hard, if not impossible, to include human knowledge to data augmentation for time series data. For example, rotating an image will not change its content or the label. While permuting a time series instance may ruin its signal patterns and generates a meaningless time series instance. In addition, the tremendous heterogeneity of real-life time series datasets further makes selections based on trial-and-errors impractical. Although multiple data augmentation methods have been proposed for time series data, there is less discussion on what is a good augmentation that is meaningful for a given learning task and dataset without prefabricated human priors. From our perspective, ideal data augmentations for contrastive representation should keep high fidelity, high variety, and adaptive to different datasets. [The illustration and examples are shown in Figure 2.](#)

High Fidelity. Augmentations with high fidelity maintain the semantic identity that is invariant to transformations. Considering the inexplicability in practical time series data, it is challenging to visually check the fidelity of augmentations. Thus, we assume that the semantic identity of a time series instance is presented by its label in the downstream task, which might be either available or unavailable during the training period. Here, we start our analysis from the supervised case and will extend it to the unsupervised case later. Based on the information theory (Tishby et al., 2000), we define the objective that keeps high fidelity as the large mutual information (MI) between augmentation v and the label y , i.e., $MI(v; y)$.

We consider augmentation v as a *probabilistic* function of x and a random variable ϵ , that $v = g(x; \epsilon)$. From the definition of mutual information, we have $MI(v; y) = H(y) - H(y|v)$, where $H(y)$ is the (Shannon) entropy of y and $H(y|v)$ is the entropy of y conditioned on augmentation v . Since $H(y)$ is irrelevant to data augmentations, the objective is equivalent to minimizing the conditional entropy $H(y|v)$. Considering the efficient optimization, we follow Ying et al. (2019) and Luo et al. (2020) to approximate it with cross-entropy between y and \hat{y} , where \hat{y} is the prediction with augmentation v as the input and calculated via

$$v = g(x; \epsilon) \quad \mathbf{z} = f_{\theta}(v) \quad \hat{y} = h_w(\mathbf{z}), \quad (1)$$

where \mathbf{z} is the representation and $h_w(\cdot)$ is a prediction projector parameterized by w . The prediction projector is optimized by the classification objective. Then, the objective of high fidelity for

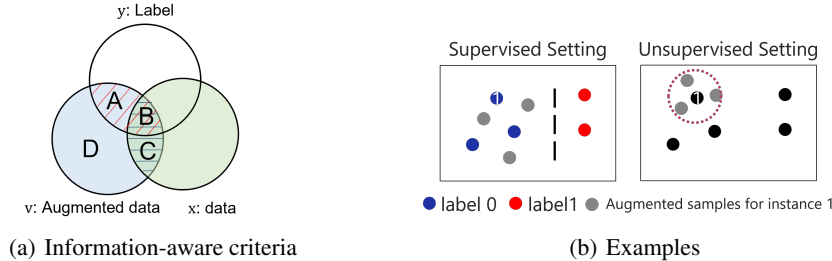


Figure 2: Illustration of the criteria. (a) The proposed criteria have two components: high fidelity, and variety. Fidelity is represented by the area of $A+B$, the mutual information between augmented data v and label y . Variety is denoted by $A+D$, the entropy of v conditioned on the raw input x . (b) In the supervised setting, good data augmentations generate instances in the area constrained by the label to enlarge the input training space. In the unsupervised setting, with one-hot-based pseudo labels, the generated instances are constrained to the region around the raw input. Such that they are still distinguishable from other instances.

supervised or semi-supervised cases is to minimize

$$\text{CE}(y; \hat{y}) = - \sum_{c=1}^C P(y = c) \log P(\hat{y} = c), \quad (2)$$

where C is the number of labels.

In the *unsupervised* settings where y is unavailable, *one-hot* encoding $y_s \in \mathbb{R}^{|\mathbb{X}|}$ is utilized as the pseudo label to replace y in Eq. (2). The motivation is that augmented instances are still distinguishable from other instances with the classifier. We theoretically show that augmentations that preserving pseudo labels have the following properties.

Property 1 (Preserving Fidelity). *If augmentation v preserves the one-hot encoding pseudo label, the mutual information between v and the downstream task label y (although not visible to training) is equivalent to that between raw input x and y , i.e., $MI(v; y) = MI(x; y)$.*

Property 2 (Adding New Information). *By preserving the one-hot encoding pseudo label, augmentation v contains new information comparing to the raw input x , i.e., $H(v) \geq H(x)$.*

Detailed proofs are shown in the Appendix A. These properties show that in the unsupervised setting, preserving the one-hot encoding pseudo label guarantees that the generated augmentations will not decrease the fidelity, regardless of the downstream tasks and variances inherent in the augmentations. Concurrently, it may introduce new information for contrastive learning.

Since the number of labels is equal to the number of instances in dataset \mathbb{X} in an unsupervised case, direct optimization of Eq. (2) is inefficient and unscalable. Thus, we further relax it by approximating y with the batch-wise one-hot encoding y_B , which decreases the number of labels C from the dataset size to the batch size.

High Variety. Sufficient variances in augmentations improve the generalization capacity of contrastive learning models. In the information theory, the uncertainty inherent in the random variable’s possible outcomes is described by its entropy. Considering that augmented instances are generated based on the raw input x , we maximize the entropy of v conditioned on x , $H(v|x)$, to maintain a high variety of augmentations. From the definition of conditional entropy, we have

$$H(v|x) = H(v) - MI(v; x). \quad (3)$$

We dismiss the first part since the unconstrained entropy of v can be dominated by meaningless noise. Considering the continuity of both v and x , we adopt a mutual information neural estimator, InfoNCE (Oord et al., 2018; Tian et al., 2020) to approximately compute the mutual information for its practical effectiveness. Other MI estimators, such as Jensen-Shannon (JSD) estimator (Nowozin et al., 2016), normalized temperature-scaled cross-entropy (NT-Xent) (Chen et al., 2020), and leave-one-out (Poole et al., 2019) can also conveniently be the plug-and-play component in our framework. Then, the objective to encourage high variety is to maximize the InfoNCE between v and x :

$$\text{InfoNCE}(v; x) = -\mathbb{E}_x \left[\log \frac{\exp(\text{sim}(\mathbf{z}_x, \mathbf{z}_v))}{\sum_{x' \in \mathbb{X}} \exp(\text{sim}(\mathbf{z}_x, \mathbf{z}_{v'}))} \right], \quad (4)$$

where v' is an augmented instance of input instance x' . \mathbf{z}_x , \mathbf{z}_v , and $\mathbf{z}_{v'}$ are representations of instance x , v , and v' respectively. $\text{sim}(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{z}_1^T \mathbf{z}_2$ is the inner product of vectors \mathbf{z}_1 and \mathbf{z}_2 .

Criteria. Combining the information aware definition of both high fidelity and variety, we propose the criteria for selecting good augmentations without prior knowledge,

$$\min_v -\text{InfoNCE}(v; x) + \beta \text{CE}(y; h_w(f_\theta(v))), \quad (5)$$

where β is a hyper-parameter to achieve the trade-off between fidelity and variety. Note that in the *unsupervised* settings, y is replaced by *one-hot* encoding pseudo label.

Relation to Information Bottleneck. Although the formation is similar to information bottleneck in data compression, $\min_{p(e|x)} \text{MI}(x; e) - \beta \text{MI}(e; y)$, our criteria are different in the following aspects. First, e in the information bottleneck is a representation of input x , while v in Eq.(5) represents the augmented instances. Second, information bottleneck aims to keep minimal and sufficient information for data compression, while our criteria are designed for data augmentations in contrastive learning. Third, in information bottleneck, the compressed representation e is a deterministic function of input x with no variances. $\text{MI}(e; y)$ and $\text{MI}(e; x)$ are constraint by $\text{MI}(x; y)$ and $H(x)$ that $\text{MI}(e; y) \leq \text{MI}(x; y)$ and $\text{MI}(e; x) = H(e)$, where $H(e)$ is the entropy of e . In our criteria, v is a probabilistic function of input x . As a result, the variances of v makes the augmentation space much larger than the compression representation space in information bottleneck.

Relation to InfoMin. InfoMin is designed based on the information bottleneck that good views should keep minimal and sufficient information from the original input (Tian et al., 2020). Similar to the information bottleneck, InfoMin assumes that augmented views are functions of the input, which heavily constrains the variance of data augmentations. Besides, high fidelity property is dismissed in the unsupervised setting. It works for image datasets due to the availability of human knowledge. However, it may fail to generate reasonable augmentations for time series data.

2.3 TIME SERIES META-CONTRASTIVE LEARNING

We aim to design a learnable augmentation selector that learns to select feasible augmentations in a data-driven manner. With such adaptive data augmentations, the contrastive loss is then used to train the encoder that learns representations from raw time series.

2.3.1 ARCHITECTURE

The adopted encoder $f_\theta(x) : \mathbb{R}^{T \times F} \rightarrow \mathbb{R}^D$ consists of two components, a fully connected layer, and a 10-layer dilated CNN module (Franceschi et al., 2019; Yue et al., 2021). To explore the inherent structure of time series, we include both global-wise (instance-level) and local-wise (subsequence-level) losses in the contrastive learning framework to train the encoder.

Global-wise contrastive loss is designed to capture the instance level relations in a time series dataset. Formally, given a batch of time series instances $\mathbb{X}_B \subseteq \mathbb{X}$, for each instance $x \in \mathbb{X}_B$, we generate an augmented instance v with an adaptively selected transformation, which will be introduced later. (x, v) is regarded as a positive pair and other $(B-1)$ combinations $\{(x, v')\}$, where v' is an augmented instance of x' and $x' \neq x$, are considered as negative pairs. Following Chen et al. (2020); You et al. (2020), we design the global-wise contrastive loss based on InfoNCE (Hjelm et al., 2018). The batch-wise instance-level contrastive loss is

$$\mathcal{L}_g = -\frac{1}{|\mathbb{X}_B|} \sum_{x \in \mathbb{X}_B} \log \frac{\exp(\text{sim}(\mathbf{z}_x, \mathbf{z}_v))}{\sum_{x' \in \mathbb{X}_B} \exp(\text{sim}(\mathbf{z}_x, \mathbf{z}_{v'}))}. \quad (6)$$

Local-wise contrastive loss is proposed to explore the intra-temporal relations in time series. For an augmented instance v of a time series instance x , we first split it into a set of subsequences \mathbb{S} , each with length L . For each subsequence $s \in \mathbb{S}$, we follow Tonekaboni et al. (2021) to generate a positive pair (s, p) by selecting another subsequence close to it. Non-neighboring samples, $\tilde{\mathcal{N}}_s$, are adopted to generate negative pairs. [Detailed descriptions can be found in Appendix B](#). Then, the local-wise contrastive loss for an instance x is:

$$\mathcal{L}_{c_x} = -\frac{1}{|\mathbb{S}|} \sum_{s \in \mathbb{S}} \log \frac{\exp(\text{sim}(\mathbf{z}_s, \mathbf{z}_p))}{\exp(\text{sim}(\mathbf{z}_s, \mathbf{z}_p)) + \sum_{j \in \tilde{\mathcal{N}}_s} \exp(\text{sim}(\mathbf{z}_s, \mathbf{z}_j))}. \quad (7)$$

Across all instances in a batch, we have $\mathcal{L}_c = \frac{1}{|\mathbb{X}_B|} \sum_{x \in \mathbb{X}_B} \mathcal{L}_{c,x}$. Then, the final contrastive objective is:

$$\min_{\theta} \mathcal{L}_g + \alpha \mathcal{L}_c, \quad (8)$$

where α is a hyper-parameter to achieve the trade-off between global and local contrastive losses.

2.3.2 META-LEARNER NETWORK

Previous time series contrastive learning methods (Franceschi et al., 2019; Fan et al., 2020; Eldele et al., 2021; Tonekaboni et al., 2021) generate augmentations with either rule of thumb guided by prefabricated human priors or tedious trial-and-errors, which are designed for specific datasets and learning tasks. In this part, we discuss how to adaptively select the optimal augmentations with a meta-learner network based on the proposed information-aware criteria. We can regard its choice of optimal augmentation as a kind of prior selection. We first choose a set of candidate transformations \mathbb{T} , such as jittering and time warping. Each candidate transformation $t_i \in \mathbb{T}$ is associated with an important weight $p_i \in (0, 1)$, inferring the probability of selecting transformation t_i . For an instance x , the augmented instance v_i through transformation t_i can be computed by:

$$a_i \sim \text{Bernoulli}(p_i) \quad v_i = (1 - a_i)x + a_i t_i(x). \quad (9)$$

Considering multiple transformations, we pad all v_i to be with the same length. Then, the adaptive augmented instance can be achieved by combing candidate ones, $v = \frac{1}{|\mathbb{T}|} \sum_i v_i$.

To enable the efficient optimization with gradient-based methods, we approximate discrete Bernoulli processes with binary concrete distributions (Maddison et al., 2016). Specifically, we approximate a_i in Eq. (9) with

$$\epsilon \sim \text{Uniform}(0, 1) \quad a_i = \sigma((\log \epsilon - \log(1 - \epsilon) + \log \frac{p_i}{1 - p_i})/\tau), \quad (10)$$

where $\sigma(\cdot)$ is the sigmoid function and τ is the temperature controlling the approximation. The rationality of such approximation is given in Appendix A. Moreover, with temperature $\tau > 0$, the gradient $\frac{\partial v}{\partial p_i}$ is well-defined. Therefore, our meta-network is end-to-end differentiable. Detailed algorithm is shown in Appendix B.

3 RELATED WORK

3.1 CONTRASTIVE TIME SERIES REPRESENTATION LEARNING

Contrastive learning has been utilized widely in representation learning with superior performances in various domains (Chen et al., 2020; Xie et al., 2019; You et al., 2020). Recently, some efforts have been devoted to applying contrastive learning to the time series domain (Oord et al., 2018; Franceschi et al., 2019; Fan et al., 2020; Eldele et al., 2021; Tonekaboni et al., 2021; Yue et al., 2021). Time Contrastive Learning trains a feature extractor with a multinomial logistic regression classifier to discriminate all segments in a time series (Hyvarinen & Morioka, 2016). Franceschi et al. (2019) generates positive and negative pairs based on subsequences. TNC employs a debiased contrastive objective to ensure that in the representation space, signals in the local neighborhood are distinguishable from non-neighboring signals (Tonekaboni et al., 2021). SelfTime adopts multiple hand-crafted augmentations for unsupervised time series contrastive learning by exploring both inter-sample and intra-sample relations (Fan et al., 2020). TS2Vec learns a representation for each time stamp and conducts contrastive learning in a hierarchical way (Yue et al., 2021). However, data augmentations in these methods are either universal or selected by error-and-trial, hindering them away from been widely applied in complex real-life datasets.

3.2 TIME SERIES FORECASTING

Forecasting is a critical task in time series analysis. Deep learning architectures used in the literature include Recurrent Neural Networks (RNNs) (Salinas et al., 2020; Oreshkin et al., 2019), Convolutional Neural Networks (CNNs) (Bai et al., 2018), Transformers (Li et al., 2019; Zhou et al., 2021), and Graph Neural Networks (GNNs) (Cao et al., 2021). N-BEATS deeply stacks fully-connected layers with backward and forward residual links for univariate times series forecasting (Oreshkin

et al., 2019). TCN utilizes a deep CNN architecture with dilated causal convolutions (Bai et al., 2018). Considering both long-term dependencies and short-term trends in multivariate time series, LSTnet combines both CNNs and RNNs in a unified model (Lai et al., 2018). LogTrans brings the Transformer model to time series forecasting with causal convolution in its attention mechanism Li et al. (2019). Informer further proposes a sparse self-attention mechanism to reduce the time complexity and memory usage (Zhou et al., 2021). StemGNN is a GNN based model that considers the intra-temporal and inter-series correlations simultaneously Cao et al. (2021). Unlike these works, we aim to learn general representations for time series data that can not only be used for forecasting but also other tasks, such as classification. Besides, the proposed framework is compatible with various architectures as encoders.

3.3 ADAPTIVE DATA AUGMENTATION

Data augmentation is an important component in contrastive learning. Existing researches reveal that the choices of optimal augmentation are dependent on downstream tasks and datasets (Chen et al., 2020; Fan et al., 2020). Some researchers have explored adaptive selections of optimal augmentations for contrastive learning in the vision field. AutoAugment automatically searches the combination of translation policies via a reinforcement learning method (Cubuk et al., 2019). Faster-AA improves the searching pipeline for data augmentation using a differentiable policy network (Hataya et al., 2020). DADA further introduces an unbiased gradient estimator for an efficient one-pass optimization strategy (Li et al., 2020). Within contrastive learning frameworks, Tian et al. (2020) applies the Information Bottleneck theory that optimal views should share minimal and sufficient information, to guide the selection of good views for contrastive learning in the vision domain. Considering the inexplicability of time series data, directly applying the InfoMin framework may keep insufficient information during augmentation. Different from Tian et al. (2020), we focus on the time series domain and propose an end-to-end differentiable method to automatically select the optimal augmentations for each dataset.

4 EXPERIMENTS

In this section, we compare InfoTS with SOTA baselines on time series forecasting and classification tasks. We also conduct case studies to show insights into the proposed criteria and meta-learning framework. Detailed experimental setups are shown in Appendix C. Full experimental results and extra experiments, such as parameter sensitivity studies, are presented in Appendix D.

4.1 TIME SERIES FORECASTING

Time series forecasting aims to predict the future L_y time stamps, with the last L_x observations. We follow Yue et al. (2021) to train a linear model regularized with the L2 norm penalty to make predictions. The output has dimension L_y in the univariate case and $L_y \times F$ for the multivariate case, where F is the feature dimension.

Datasets and Baselines. Four benchmark datasets for time series forecasting are adopted, including ETTh1, ETTh2, ETTm1 (Zhou et al., 2021), and the Electricity dataset (Dua & Graff, 2017). These datasets are used in both univariate and multivariate settings. We compare unsupervised InfoTS to the SOTA baselines, including TS2Vec (Yue et al., 2021), Informer (Zhou et al., 2021), StemGNN (Cao et al., 2021), TCN (Bai et al., 2018), LogTrans (Li et al., 2019), LSTnet (Lai et al., 2018), and N-BEATS (Oreshkin et al., 2019). Among these methods, N-BEATS is merely designed for the univariate and StemGNN is for multivariate only. We refer to Yue et al. (2021) to set up baselines for a fair comparison. Standard metrics for a regression problem, Mean Squared Error (MSE), and Mean Absolute Error (MAE) are utilized for evaluation. Evaluation results of univariate time series forecasting are shown in Table 1, while multivariate forecasting results are reported in the Appendix D.4 due to the space limitation.

Performance. As shown in Tabel 1 and Tabel 4, comparison in both univariate and multivariate settings indicates that InfoTS consistently matches or outperforms the leading baselines. Some results of StemGNN are unavailable due to the out-of-memory issue (Yue et al., 2021). Specifically, we have the following observations. **TS2Vec, another contrastive learning method with data augmentations, achieves the second-best performance in most cases. The consistent improvement of TS2Vec**

Table 1: Univariate time series forecasting results.

Dataset	L_y	InfoTS		TS2Vec		Informer		LogTrans		N-BEATS		TCN		LSTnet	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh ₁	24	0.039	0.148	0.039	0.152	0.098	0.247	0.103	0.259	0.094	0.238	0.075	0.210	0.108	0.284
	48	0.056	0.179	0.062	0.191	0.158	0.319	0.167	0.328	0.210	0.367	0.227	0.402	0.175	0.424
	168	0.099	0.237	0.134	0.282	0.183	0.346	0.207	0.375	0.232	0.391	0.316	0.493	0.396	0.504
	336	0.114	0.260	0.154	0.310	0.222	0.387	0.230	0.398	0.232	0.388	0.306	0.495	0.468	0.593
	720	0.147	0.309	0.163	0.327	0.269	0.435	0.273	0.463	0.322	0.490	0.390	0.557	0.659	0.766
ETTh ₂	24	0.081	0.215	0.090	0.229	0.093	0.240	0.102	0.255	0.198	0.345	0.103	0.249	3.554	0.445
	48	0.115	0.261	0.124	0.273	0.155	0.314	0.169	0.348	0.234	0.386	0.142	0.290	3.190	0.474
	168	0.171	0.327	0.208	0.360	0.232	0.389	0.246	0.422	0.331	0.453	0.227	0.376	2.800	0.595
	336	0.183	0.341	0.213	0.369	0.263	0.417	0.267	0.437	0.431	0.508	0.296	0.430	2.753	0.738
	720	0.194	0.357	0.214	0.374	0.277	0.431	0.303	0.493	0.437	0.517	0.325	0.463	2.878	1.044
ETTM ₁	24	0.014	0.085	0.015	0.092	0.030	0.137	0.065	0.202	0.054	0.184	0.041	0.157	0.090	0.206
	48	0.024	0.113	0.027	0.126	0.069	0.203	0.078	0.220	0.190	0.361	0.101	0.257	0.179	0.306
	96	0.036	0.142	0.044	0.161	0.194	0.372	0.199	0.386	0.183	0.353	0.142	0.311	0.272	0.399
	288	0.077	0.208	0.103	0.246	0.401	0.554	0.411	0.572	0.186	0.362	0.318	0.472	0.462	0.558
	672	0.115	0.257	0.156	0.307	0.512	0.644	0.598	0.702	0.197	0.368	0.397	0.547	0.639	0.697
Electricity	24	0.249	0.273	0.260	0.288	0.251	0.275	0.528	0.447	0.427	0.330	0.263	0.279	0.281	0.287
	48	0.294	0.299	0.319	0.324	0.346	0.339	0.409	0.414	0.551	0.392	0.373	0.344	0.381	0.366
	168	0.401	0.368	0.427	0.394	0.544	0.424	0.959	0.612	0.893	0.538	0.609	0.462	0.599	0.500
	336	0.536	0.453	0.565	0.474	0.713	0.512	1.079	0.639	1.035	0.669	0.855	0.606	0.823	0.624
Avg.		0.155	0.254	0.175	0.278	0.263	0.367	0.336	0.419	0.338	0.402	0.289	0.359	1.090	0.516

Table 2: Multivariate time series classification on 30 UEA datasets.

Method	InfoTS _s	InfoTS	TS2Vec	T-Loss	TNC	TS-TCC	TST	DTW
Avg. ACC	0.724	0.711	0.704	0.658	0.670	0.668	0.617	0.629
Avg. Rank	2.133	2.633	2.933	3.733	4.233	4.033	4.867	4.233

over other baselines indicates the effectiveness of contrastive learning for time series representations learning. However, such universal data augmentations may not be the most informative ones to generate positive pairs. Comparing to TS2Vec, InfoTS decreases the average MSE by 11.4%, and the average MAE by 8.6% in the univariate setting. In the multivariate setting, the MSE and MAE decrease by 4.6% and 2.3%, respectively. The reason is that InfoTS can adaptively select the most suitable augmentations in a data-driven manner with high variety and high fidelity. Encoders trained with such informative augmentations learn representations with higher quality.

4.2 TIME SERIES CLASSIFICATION

Following the previous setting, we evaluate the quality of representations on time series classification in a standard supervised manner (Franceschi et al., 2019; Yue et al., 2021). We train an SVM classifier with a radial basis function kernel on top of representations in the training split and then compare the prediction in the test set.

Datasets and Baselines. Two types of benchmark datasets are used for evaluations. The UCR archive (Dau et al., 2019) is a collection of 128 univariate time series datasets and the UEA archive (Bredin, 2017) consists of 30 multivariate datasets. We compare InfoTS with baselines including TS2Vec (Yue et al., 2021), T-Loss (Franceschi et al., 2019), TS-TCC (Eldele et al., 2021), TST (Zerveas et al., 2021), and DTW (Franceschi et al., 2019). For our methods, InfoTS_s, training labels are only used to train the meta-network to select suitable augmentations, and InfoTS is with a purely unsupervised setting for representation learning.

Performance. The results on the UEA datasets are summarized in Table 2. Full results are provided in Appendix D.4. With the ground-truth label guiding the meta-network, InfoTS_s substantially outperforms other baselines. On average, it improves the classification accuracy by 2.8% over the best baseline, TS2Vec, with an average rank value 2.133 on all 30 UEA datasets. Under the purely unsupervised setting, InfoTS preserves fidelity by adopting one-hot encoding as the pseudo labels. InfoTS achieves the second best average performance in Table 2, with an average rank value 2.633. Performances on the 128 UCR datasets are shown in Table 9 in Appendix D.4. These datasets are univariate with easily recognized patterns, where data augmentations have marginal or even negative effects (Yue et al., 2021). However, with adaptively selected augmentations for each dataset based on our criteria, InfoTS_s and InfoTS still outperform the state-of-the-arts.

4.3 EVALUATION OF THE CRITERIA

In Section 2.2, we propose criteria of data augmentations for time series based on information theory that good augmentations should have high variety and fidelity. With mutual information neural

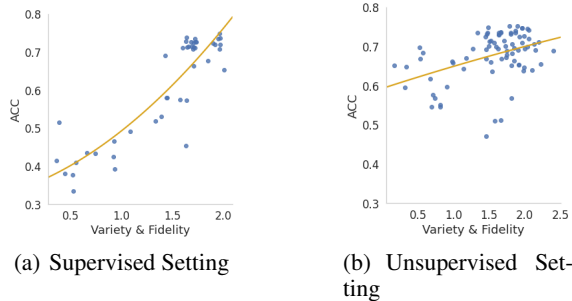


Figure 3: Evaluation of the criteria.

Table 3: Ablation studies.

L_y	InfoTS		Contrastive Objective				Data Augmentation				Meta Objective			
			w/o Local		w/o Global		Random		All		w/o Fidelity		w/o Variety	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
24	0.249	0.273	0.249	0.273	0.261	0.385	0.255	0.280	0.255	0.276	0.248	0.274	0.251	0.273
48	0.294	0.299	0.298	0.304	0.307	0.313	0.301	0.308	0.303	0.308	0.294	0.301	0.293	0.299
168	0.401	0.368	0.405	0.373	0.420	0.385	0.413	0.378	0.410	0.376	0.402	0.371	0.403	0.369
336	0.536	0.453	0.545	0.457	0.557	0.474	0.549	0.466	0.552	0.470	0.537	0.460	0.537	0.459
Avg.	0.370	0.348	0.380	0.358	0.380	0.357	0.374	0.352	0.386	0.365	0.370	0.351	0.371	0.350

estimation and cross-entropy as approximations, we get the criteria in Eq. (5). To empirically verify the effectiveness of the proposed criteria, we adopt the dataset CricketX from the UCR archive (Dau et al., 2019; Fan et al., 2020), and conduct augmentations with different configurations, such as jittering with noise from $\mathcal{N}(0, 0.3)$ and cutout 10% time stamps. For each configuration, we calculate the criteria score and the corresponding classification accuracy within the setting in Section 4.2. As shown in Figure 3, in general, accuracy performance is positively related to the proposed criteria in both supervised and unsupervised settings, [verifying the correctness of using the criteria as the objective in the meta-network training.](#)

4.4 ABLATION STUDIES

To present deep insights into the proposed method, we verify key components in InfoTS with multiple ablation studies on the Electricity dataset. Results are shown in Table 3. 1) We verify the effectiveness of the adopted contrastive learning objective, Eq. (8), by comparing InfoTS with its two variants. “w/o Local” removes the local contrastive loss between subsequences and “w/o Global” removes the instance-level contrastive loss. The comparison between these variants and InfoTS demonstrates that both instance-level and intra-temporal objectives are important in contrastive learning for time series data. 2) To demonstrate the advantage of adaptive selection of augmentations, we compare InfoTS with variants “Random” and “All”. “Random” randomly selects an augmentation from candidate transformation functions each time and “All” sequentially applies transformations to generate augmented instances. Performance decreases are observed in these variances, verifying the key role of adaptive selection in our method. 3) To show the effects of variety and fidelity objectives in meta-network training, we include two variants, “w/o Fidelity” and “w/o Variety”, which dismiss the fidelity or variety objective, respectively. The comparison between InfoTS and the variants empirically confirms both variety and fidelity are important for data augmentation in contrastive learning. [Note that the adopted basic augmentations are manually tuned in a previous paper \(Fan et al., 2020\) with high qualities for contrastive learning. Including low quality basic augmentations, we conduct more ablation studies are shown in Appendix D.3.](#)

5 CONCLUSIONS

We propose criteria of data augmentations for time series data based on the information theory that good augmentations should preserve high variety and high fidelity. We approximate the criteria with mutual information neural estimation and cross-entropy estimation. Based on the approximated criteria, we present a meta-learning method to adaptively select optimal augmentations for contrastive representation learning. Comprehensive experiments show that representations produced by our method are high qualified and easy to use in various downstream tasks, such as time series forecasting and classification, with state-of-the-art performances.

STATEMENTS

Reproducibility Statement. The training algorithm of the proposed InfoTS is shown in Appendix B. Detailed experimental settings, including architectures and hyper-parameters settings, are given in Appendix C. Core source codes with an example of usages are attached in the supplementary material. Full version will be made available upon acceptance.

REFERENCES

- Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.
- Hervé Bredin. Tristounet: triplet loss for speaker turn embedding. In *ICASSP*, pp. 5430–5434, 2017.
- Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Conguri Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. Spectral temporal graph neural network for multivariate time-series forecasting. *arXiv preprint arXiv:2103.07719*, 2021.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607, 2020.
- Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, pp. 113–123, 2019.
- Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.
- Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.
- Haoyi Fan, Fengbin Zhang, and Yue Gao. Self-supervised time series representation learning by inter-intra relational reasoning. *arXiv preprint arXiv:2011.13548*, 2020.
- Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. *arXiv preprint arXiv:1901.10738*, 2019.
- Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. Faster autoaugment: Learning augmentation strategies using backpropagation. In *ECCV*, pp. 1–16. Springer, 2020.
- Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
- Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *NIPS*, pp. 3765–3773, 2016.
- Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *SIGIR*, pp. 95–104, 2018.

- Arthur Le Guennec, Simon Malinowski, and Romain Tavenard. Data Augmentation for Time Series Classification using Convolutional Neural Networks. In *ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data*, 2016.
- Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyu Zhou, Wenhui Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *NeurIPS*, pp. 5243–5253, 2019.
- Yonggang Li, Guosheng Hu, Yongtao Wang, Timothy Hospedales, Neil M Robertson, and Yongxin Yang. Dada: Differentiable automatic data augmentation. *arXiv preprint arXiv:2003.03780*, 2020.
- Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. *arXiv preprint arXiv:2011.04573*, 2020.
- Dongsheng Luo, Wei Cheng, Jingchao Ni, Wenchao Yu, Xuchao Zhang, Bo Zong, Yanchi Liu, Zhengzhang Chen, Dongjin Song, Haifeng Chen, et al. Unsupervised document embedding via contrastive augmentation. *arXiv preprint arXiv:2103.14542*, 2021.
- Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.
- Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization, 2016.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Boris N Oreshkin, Dmitri Carпов, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.
- David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.
- Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020.
- Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.
- Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*, 2021.
- Qingsong Wen, Liang Sun, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. In *AAAI*, 2021.
- Mark van der Wilk, Matthias Bauer, ST John, and James Hensman. Learning invariances using the marginal likelihood. In *NeurIPS*, pp. 9960–9970, 2018.
- Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

- Qiang Yang and Xindong Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(04):597–604, 2006.
- Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, volume 32, pp. 9240, 2019.
- Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *NeurIPS*, pp. 5812–5823, 2020.
- Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, and Bixiong Xu. Ts2vec: Towards universal representation of time series. *arXiv preprint arXiv:2106.10466*, 2021.
- George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *SIGKDD*, pp. 2114–2124, 2021.
- Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, 2021.

A DETAILED PROOFS

Properties of data augmentations that preserve pseudo labels.

We assume that v_i and v_j are two augmented instances of inputs x_i and x_j , respectively. Preserving pseudo labels defined in one-hot encoding requires that the map between variable x and v is one-to-many. Formally, $x_i \neq x_j \rightarrow v_i \neq v_j$. This can be proved by contradiction. If we have a pair of (i, j) that $i \neq j$ and $v_i = v_j$, then we have $f_\theta(v_i) = f_\theta(v_j)$, showing that augmentations cannot preserve pseudo labels.

Property 1 (Preserving Fidelity). *If augmentation v preserves the one-hot encoding pseudo label, the mutual information between v and downstream task label y (although not visible to training) is equivalent to that between raw input x and y , i.e., $MI(v; y) = MI(x; y)$.*

Proof. From the definition of mutual information, we have

$$\begin{aligned} MI(v; y) &= H(y) - H(y|v) = H(y) + \sum_{v,y} p(v, y) \log \frac{p(v, y)}{p(v)} \\ &= H(y) + \sum_{x,y} \sum_{v \in \mathbb{V}(x)} p(v, y) \log \frac{p(v, y)}{p(v)}, \end{aligned}$$

where $\mathbb{V}(x)$ is the set of augmented instances of a time series instance x . In the unsupervised setting where the ground-truth label y is unknown, we assume that the augmentation v is a (probabilistic) function of x only. The only qualifier means $p(v|x, y) = p(v|x)$. Since the mapping from x to v is one-to-many. For each $v \in \mathbb{V}(x)$ we have $p(v, y) = p(v, x, y)$ and $p(v) = p(v|x)p(x)$. Thus, we have

$$\begin{aligned} \frac{p(v, y)}{p(v)} &= \frac{p(v, x, y)}{p(v|x)p(x)} = \frac{p(v|x, y)p(x, y)}{p(v|x)p(x)} = \frac{p(v|x)p(x, y)}{p(v|x)p(x)} = \frac{p(x, y)}{p(x)} \\ MI(v; y) &= H(y) + \sum_{x,y} \sum_{v \in \mathbb{V}(x)} p(v, y) \log \frac{p(x, y)}{p(x)} \\ &= H(y) + \sum_{x,y} \left[\sum_{v \in \mathbb{V}(x)} p(v, y) \right] \log \frac{p(x, y)}{p(x)} \\ &= H(y) + \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)} \\ &= MI(x; y). \end{aligned}$$

Property 2 (Adding New Information). *By preserving the one-hot encoding pseudo label, augmentation v contains new information comparing to the raw input x , i.e., $H(v) \geq H(x)$.*

In information theory, entropy describes the amount of information of a random variable. For simplicity, we assume a finite number of augmented instances for each input, and each augmented instance is generated independently. Then, we have $p(x) = \sum_{v \in \mathbb{V}(x)} p(v)$. Then we have that the entropy of variable x is no larger than the entropy of v .

$$\begin{aligned} H(x) &= - \sum_x p(x) \log p(x) = - \sum_x \left[\sum_{v \in \mathbb{V}(x)} p(v) \right] \log \left[\sum_{v \in \mathbb{V}(x)} p(v) \right] \\ &= - \sum_x \sum_{v \in \mathbb{V}(x)} p(v) \log \left[\sum_{v \in \mathbb{V}(x)} p(v) \right] \\ &\leq - \sum_x \sum_{v \in \mathbb{V}(x)} p(v) \log p(v) \\ &= - \sum_v p(v) \log p(v) = H(v) \end{aligned}$$

Rationality of approximation of Bernoulli distribution with binary concrete distribution in Eq. (10).

In the binary concrete distribution, parameter τ controls the temperature that achieves the trade-off between binary output and continuous optimization. When $\tau \rightarrow 0$, we have $\lim_{\tau \rightarrow 0} P(a_i = 1) = p_i$, which is equivalent to the Bernoulli distribution.

Proof.

$$\begin{aligned} \lim_{\tau \rightarrow 0} P(a_i = 1) &= \lim_{\tau \rightarrow 0} P(\sigma((\log \epsilon - \log(1 - \epsilon) + \log \frac{p_i}{1 - p_i})/\tau) = 1) \\ &= P(\log \epsilon - \log(1 - \epsilon) + \log \frac{p_i}{1 - p_i} > 0) \\ &= P(\log \frac{\epsilon}{1 - \epsilon} - \log \frac{1 - p_i}{p_i} > 0) \\ &= P(\frac{\epsilon}{1 - \epsilon} > \frac{1 - p_i}{p_i}) \end{aligned}$$

Since ϵ , and p_i are both in $(0, 1)$, and function $\frac{x}{1-x}$ are monotonically increasing in this region. Thus, we have

$$\lim_{\tau \rightarrow 0} P(a_i = 1) = P(\epsilon > 1 - p_i) = p_i$$

B ALGORITHMS

B.1 TRAINING ALGORITHM

The training algorithm of InfoTS under both supervised and unsupervised settings is described in Algorithm 1. We first randomly initiate parameters in the encoder, meta-network, and classifier (line 2). Given a batch of training instances $\mathbb{X}_B \subseteq \mathbb{X}$, for each candidate transformation t_i , we utilize binary concrete distribution to get parameters a_i (lines 6-7), which indicates whether the transformation should be applied (line 8). $\mathbb{V}_B^{(i)}$ denotes the batch of augmented instances generated from transformation function t_i . The final augmented instances are generated by adaptively considering all candidates transformations (line 10). Parameters θ in the encoder are updated by minimizing the contrastive objective (lines 11-13). Meta-network is then optimized with information-aware criteria (lines 14-16). Then, classifier h_w is optimized with the classification objective (line 17).

Algorithm 1: Algorithm for InfoTS in both supervised and unsupervised settings

- 1: **Input:** time series dataset \mathbb{X} , the label set \mathbb{Y} (supervised setting), a set of candidate transformations \mathbb{T} , hyper-parameters α and β ,
 - 2: Initialize the encoder f_θ , parameters in meta-network $\{q_i\}_{i=1}^{|\mathbb{T}|}$, and the classifier h_w .
 - 3: **for** each epoch **do**
 - 4: **for** each training batch $\mathbb{X}_B \subseteq \mathbb{X}$ **do**
 - 5: **for** each transformation $t_i \in \mathbb{T}$ **do**
 - 6: $p_i \leftarrow \sigma(q_i)$
 - 7: $a_i \leftarrow \text{binaryConcrete}(p_i)$
 - 8: $\mathbb{V}_B^{(i)} \leftarrow \text{transform each } x \in \mathbb{X}_B \text{ with Eq. (9)}$
 - 9: **end for**
 - 10: Get \mathbb{V}_B by averaging $\{\mathbb{V}_B^{(i)}\}_{i=1}^{|\mathbb{T}|}$.
 - 11: Compute global contrastive loss \mathcal{L}_g with Eq. (6)
 - 12: Compute local contrastive loss \mathcal{L}_c with Eq. (7)
 - 13: Update parameters θ in the encoder with Eq. (8)
 - 14: Compute fidelity loss with Eq. (2)
 - 15: Compute variety loss with Eq. (4)
 - 16: Update parameters $\{q_i\}_{i=1}^{|\mathbb{T}|}$ in the meta-network with Eq. (5)
 - 17: Update parameters w in the classifier h_w with the cross-entropy loss
 - 18: **end for**
 - 19: **end for**
-

B.2 IMPLEMENTATION OF LOCAL-WISE CONTRASTIVE

Local-wise contrastive loss aims to capture the intra-temporal relations in each time series instance. For an augmented instance v , we first split it into multiple subsequences, as shown in Figure 4.

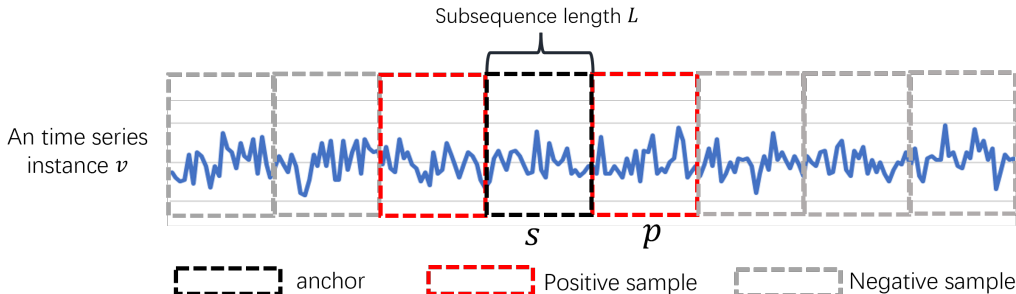


Figure 4: Positive and negative samples for a subsequence s .

Each subsequence has length L . For each subsequence s , the neighboring subsequences within window size 1 are considered as positive samples. If s locates at the end of v , then we choose the subsequence in front of s as the positive pair p . Otherwise, we choose the subsequence following s instead. Subsequences out of window size 1 are considered as negative samples.

C EXPERIMENTAL SETTINGS

C.1 DATA AUGMENTATIONS

We follow Fan et al. (2020) to set up candidate data augmentations, including jittering, scaling, cutout, time warping, window slicing, window warping and subsequence augmentation (Franceschi et al., 2019). Detailed descriptions are listed as follows.

- *Jittering* augmentation adds the random noise sampled from a Gaussian distribution $\mathcal{N}(0, 0.3)$ to the input time series.
- *Scaling* augmentation multiplies the input time series by a scaling factor sampled from a Gaussian distribution $\mathcal{N}(0, 0.5)$.
- *Cutout* operation replaces features of 10% randomly sampled time stamps of the input with zeros.
- *Time warping* random changes the speed of the timeline¹. The number of speed changes is 100 and the maximal ratio of max/min speed is 10. **If necessary, over-sampling or sampling methods are adopted to ensure the length of the augmented instance is the same as the original one.**
- *Window slicing* randomly crops half the input time series and then linearly interpolates it back to the original length (Le Guennec et al., 2016).
- *Window warping* first randomly selects 30% of the input time series along the timeline and then warps the time dimension by 0.5 or 2. Finally, we adopt linear interpolation to transform it back to the original length (Le Guennec et al., 2016).
- *Subsequence* operation random selects a subsequence from the input time series (Yue et al., 2021).

With the first 100 time stamps in the univariate Electricity dataset as an example, we visualize the original time series and the augmented ones in Figure 5.

C.2 HARDWARE AND IMPLEMENTATIONS

All experiments are conducted on a Linux machine with 4 NVIDIA GeForce RTX 2080 Ti GPUs, each with 11GB memory. CUDA version is 10.1 and Driver Version is 418.56. Our method InfoTS is implemented with Python 3.7.7 and Pytorch 1.7.1.

¹<https://tsaug.readthedocs.io/>

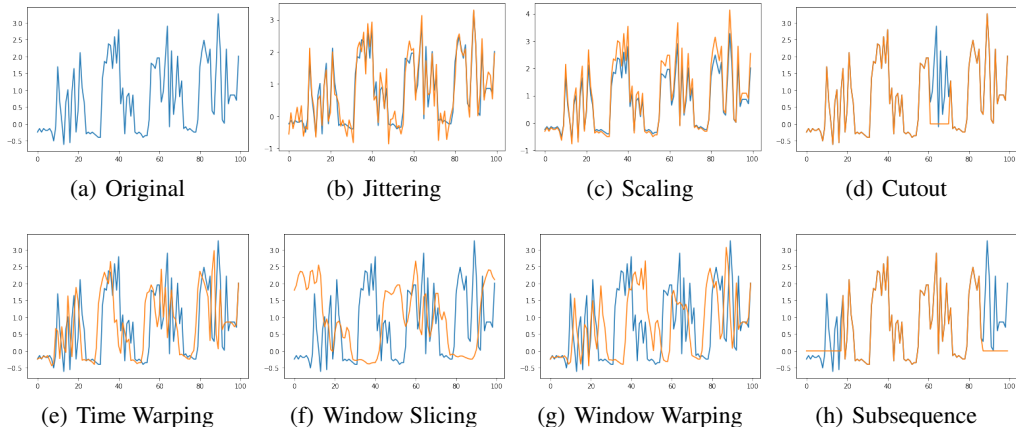


Figure 5: Examples of candidate augmentations on Electricity univariate dataset. Blue lines are the original time series data and orange ones are augmented instances.

C.3 HYPERPARAMETERS

We train and evaluate our methods with the following hyperparameters and configurations.

- **Optimizer:** Adam optimizer (Kingma & Ba, 2014) with learning rate and decay rates setting to 0.001 and (0.9,0.999), respectively.
- **SVM:** scikit-learn implementation (Pedregosa et al., 2011) with penalty $C \in \{10^i | i \in [-4, 4] \cup \infty\}$ Franceschi et al. (2019).
- **Encoder architecture:** We follow Yue et al. (2021) to design the encoder. Specifically, the output dimension of the linear projection layer is set to 64, the same for the number of channels in the following dilated CNN module. In the CNN module, GELU (Hendrycks & Gimpel, 2016) is adopted as the activation function, and the kernel size is set to 3. The dilation is set to 2^i in the i -th block.
- **Classifier architecture:** a fully connected layer that maps the representations to the label is adopted.
- **Trade-off hyperparameters:** β in Eq. (5) and α in Eq. (8) are searched in $[0.1, 0.5, 1.0, 5, 10]$. Parameter sensitivity studies are shown in Section D.1.
- **Temperature in binary concrete distribution:** we follow the practice in (Jang et al., 2016) to adopt the strategy by starting the training with a high temperature and anneal to a small value with a guided schedule.

D MORE EXPERIMENTAL RESULTS

D.1 PARAMETER SENSITIVITY STUDIES

In this part, we adopt the electricity dataset to analyze the effects of two important hyper-parameters in our method InfoTS. The hyperparameter α in Eq. (8) controls the trade-off between local and global contrastive losses when training the encoder. β in Eq. (5) achieves the balance between high variety and high fidelity when training the meta-network. We tune these parameters in range $[0.1, 0.5, 1.0, 5, 10]$ and show the results in Figure 6. These figures show that our method achieves high performance with a wide range of selections, demonstrating the robustness of the proposed method. In general, setting trade-off parameters to 0.5 or 1 achieves good performance.

D.2 CASE STUDY ON SIGNAL DETECTION

In this part, we show the potential usage of InfoTS to detect the informative signals in the time series. We adopt the CricketX dataset as an example for the case study. Subsequence augmentations

Table 4: Multivariate time series forecasting results.

Dataset	L_y	InfoTS		TS2Vec		Informer		StemGNN		TCN		LogTrans		LSTnet	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh ₁	24	0.564	0.520	0.599	0.534	0.577	0.549	0.614	0.571	0.767	0.612	0.686	0.604	1.293	0.901
	48	0.607	0.553	0.629	0.555	0.685	0.625	0.748	0.618	0.713	0.617	0.766	0.757	1.456	0.960
	168	0.746	0.638	0.755	0.636	0.931	0.752	0.663	0.608	0.995	0.738	1.002	0.846	1.997	1.214
	336	0.904	0.722	0.907	0.717	1.128	0.873	0.927	0.730	1.175	0.800	1.362	0.952	2.655	1.369
	720	1.098	0.811	1.048	0.790	1.215	0.896	–	–	1.453	1.311	1.397	1.291	2.143	1.380
ETTh ₂	24	0.397	0.456	0.398	0.461	0.720	0.665	1.292	0.883	1.365	0.888	0.828	0.750	2.742	1.457
	48	0.584	0.572	0.580	0.573	1.457	1.001	1.099	0.847	1.395	0.960	1.806	1.034	3.567	1.687
	168	1.787	1.035	1.901	1.065	3.489	1.515	2.282	1.228	3.166	1.407	4.070	1.681	3.242	2.513
	336	2.249	1.187	2.304	1.215	2.723	1.340	3.086	1.351	3.256	1.481	3.875	1.763	2.544	2.591
	720	2.511	1.316	2.650	1.373	3.467	1.473	–	–	3.690	1.588	3.913	1.552	4.625	3.709
ETTm ₁	24	0.391	0.408	0.443	0.436	0.323	0.369	0.620	0.570	0.324	0.374	0.419	0.412	1.968	1.170
	48	0.503	0.475	0.582	0.515	0.494	0.503	0.744	0.628	0.477	0.450	0.507	0.583	1.999	1.215
	96	0.537	0.503	0.622	0.549	0.678	0.614	0.709	0.624	0.636	0.602	0.768	0.792	2.762	1.542
	288	0.653	0.579	0.709	0.609	1.056	0.786	0.843	0.683	1.270	1.351	1.462	1.320	1.257	2.076
	672	0.757	0.642	0.786	0.655	1.192	0.926	–	–	1.381	1.467	1.669	1.461	1.917	2.941
Electricity	24	0.255	0.350	0.287	0.374	0.312	0.387	0.439	0.388	0.305	0.384	0.297	0.374	0.356	0.419
	48	0.279	0.368	0.307	0.388	0.392	0.431	0.413	0.455	0.317	0.392	0.316	0.389	0.429	0.456
	168	0.302	0.385	0.332	0.407	0.515	0.509	0.506	0.518	0.358	0.423	0.426	0.466	0.372	0.425
	336	0.320	0.399	0.349	0.420	0.759	0.625	0.647	0.596	0.349	0.416	0.365	0.417	0.352	0.409
Avg.		0.813	0.627	0.852	0.645	1.164	0.781	0.977	0.706	1.243	0.854	1.402	1.032	1.836	1.374

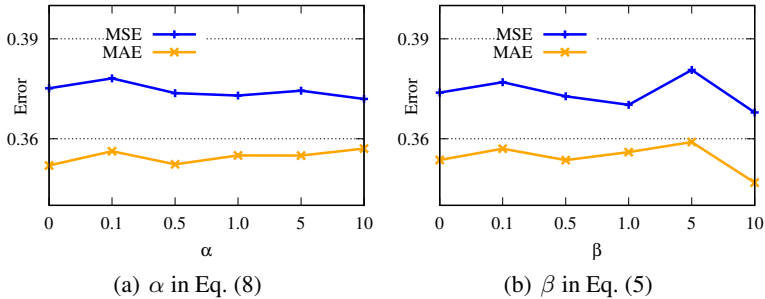


Figure 6: Parameter sensitivity studies.

on 0-100, 100-200, and 200-300 periods are adopted as candidate transformations. We observe that the one operated on 100-200 period has high fidelity and variety, leading to better accuracy performance, which is consistent with the visualization results in Figure 7.

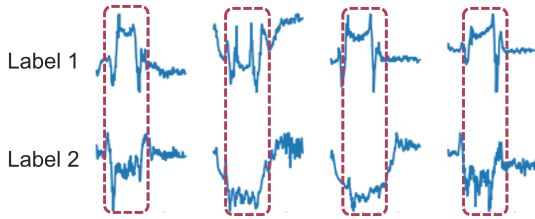


Figure 7: The informative signals locate in the middle periods of time series in the CricketX dataset.

D.3 MORE ABLATION STUDIES

D.3.1 UPDATING PROCESS OF INFOTS

To show that our InfoTS can adaptively detect the most effective augmentation based on the data distribution, we follow the setting in Section 4.4 and conduct more ablation studies to investigate comprehensively into the proposed model. We compare performances of variants that each applies a single transformation to generate augmented instances in Table 5. From the table, we know that augmentation with subsequence benefits the most for the Electricity dataset. We visualize the weight

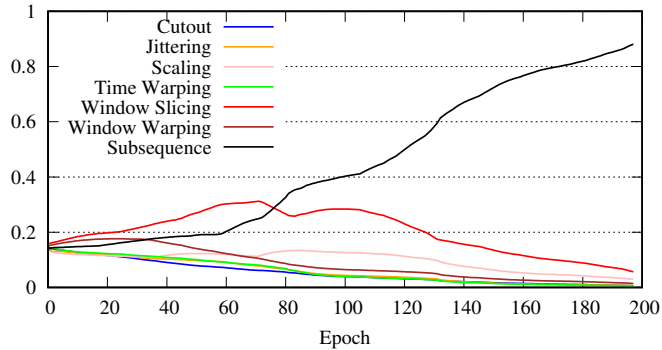


Figure 8: Weight updating process of meta-network in InfoTS.

Table 5: Effectiveness of each candidate data transformation.

L_y	InfoTS		Cutout		Jittering		Scaling		Time Warp		Window Slice		Window Warp		Subsequence	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
24	0.249	0.273	0.254	0.277	0.251	0.275	0.252	0.273	0.251	0.274	0.258	0.280	0.253	0.277	0.248	0.273
48	0.294	0.299	0.304	0.309	0.297	0.302	0.302	0.307	0.305	0.309	0.310	0.314	0.307	0.310	0.295	0.301
168	0.401	0.368	0.412	0.381	0.4.3	0.373	0.407	0.377	0.415	0.382	0.415	0.381	0.416	0.382	0.405	0.372
336	0.536	0.453	0.555	0.465	0.545	0.458	0.552	0.461	0.555	0.469	0.551	0.470	0.554	466	0.546	0.456
Avg.	0.370	0.348	0.381	0.358	0.374	0.352	0.377	0.354	0.381	0.359	0.383	0.361	0.383	0.359	0.374	0.350

updating process of InfoTS in Figure 8, with each line representing the **normalized** importance score of the corresponding transformation. The weight for subsequence increase with the epoch, showing that InfoTS tends to adopt subsequence as the optimal transformation. Consistency between accuracy performance and weight updating process demonstrates the effectiveness of InfoTS to adaptively select feasible transformations. Besides, as shown in Table 5, InfoTS outperforms the variant that uses subsequence only. This comparison shows that the meta-network learns to consider the combinations, which is better than any (single) candidate augmentation.

D.3.2 INFOTS WITH CONTROLLABLE AUGMENTATIONS.

We observe that in Table 3, the improvement of InfoTS over its two variants without Variety and Fidelity objectives is not significant. The reason is that basic augmentations are adopted from a previous paper (Fan et al., 2020), which have been manually tuned to generate relatively good performances for contrastive learning. To provide deep insights into the proposed criteria and model, in this part, we consider more augmentations with controllable hyper-parameters for the Electricity dataset.

Two groups of augmentations, subsequence augmentations with different lengths and jitter augmentations with different standard deviations, are considered in this part. Subsequence augmentations work on the temporal dimension, and jitter augmentations work on the feature dimension. For the subsequence augmentations, we range the ratio of subsequences r in the range $[0.01, 0.99]$. The subsequence augmentation with ratio r is denoted by Sub_r , such as $\text{Sub}_{0.01}$. For the jitter augmentations, the standard deviations are chosen from the range $[0.01, 3.0]$. The jitter augmentation with standard deviation std is denoted by Jitter_{std} , such as $\text{Jitter}_{0.01}$.

Intuitively, with r increasing, Sub_r generates augmented instances with lower variety and higher fidelity. For example, with $r = 0.01$, Sub_r generates subsequences that only keep 1% time stamps from the original input, leading to high variety but extremely low fidelity. Similarly, for jitter augmentations, with std increasing, Jitter_{std} generates augmented instances with higher variety but lower fidelity. We show their forecasting performances in Table 6 and Table 7. We observed that augmentations that only keeps high variety, such as $\text{Sub}_{0.01}$, $\text{Jitter}_{3.0}$ achieve worse performances in terms of MSE and MAE compared to the ones considering both criteria, such as $\text{Sub}_{0.3}$ and $\text{Jitter}_{0.3}$. On the other hand, augmentations that only keep high fidelity, such as $\text{Sub}_{0.99}$, $\text{Jitter}_{0.01}$, and $\text{Jitter}_{0.1}$ cannot lead to satisfactory performances either. Thus, we empirically verify the importance of both variety and fidelity for the selection of optimal data augmentations for contrastive learning. In Fig-

Table 6: Forecasting performances of Sub_{τ} .

L_y	0.01		0.3		0.5		0.7		0.99	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
24	0.262	0.292	0.249	0.278	0.278	0.277	0.252	0.279	0.253	0.279
48	0.309	0.321	0.300	0.310	0.298	0.306	0.302	0.309	0.305	0.311
168	0.417	0.395	0.310	0.384	0.409	0.378	0.418	0.384	0.414	0.381
336	0.550	0.475	0.550	0.467	0.553	0.460	0.562	469	0.558	0.463
Avg.	0.384	0.371	0.379	0.360	0.377	0.355	0.384	0.362	0.383	0.358

Table 7: Forecasting performances of $Jitter_{std}$.

L_y	0.01		0.1		0.3		1.0		3.0	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
24	0.252	0.277	0.253	0.278	0.250	0.276	0.249	0.250	0.281	0.277
48	0.303	0.314	0.303	0.310	0.295	0.304	0.300	0.302	0.314	0.306
168	0.416	0.385	0.413	0.383	0.401	0.373	0.418	0.418	0.391	0.378
336	0.559	0.473	0.554	0.472	0.545	0.458	0.559	0.570	0.476	0.460
Avg.	0.383	0.362	0.381	0.361	0.373	0.353	0.381	0.366	0.385	0.366

ure 9, we show the relationship between forecasting performance and our proposed criteria. The results are consistent with the conclusion drawn from classification results that forecasting performance, evaluated with both MSE and MAE, is positively related to the proposed criteria.

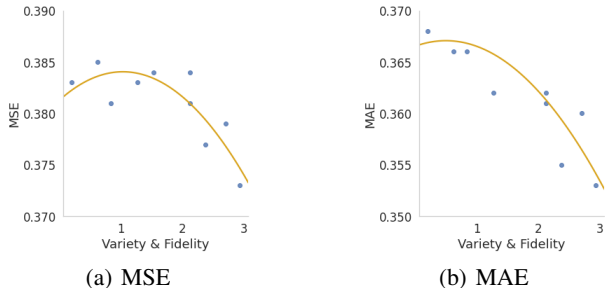


Figure 9: Evaluation of the criteria on forecasting.

With the proposed criteria, we empirically show the advantage of the developed meta-network on learning optimal augmentations. The overall observations are consistent with ones from Section 4.4. Here, we provide more details. 1) “Random” is a variant that randomly selects a candidate augmentation method to generate augmented instances. “All” combines all candidate augmentations evenly to generate augmented instances. These two variances are routinely utilized in previous works (Chen et al., 2020; You et al., 2020). They work well when candidate augmentations are manually verified. Otherwise, their performances are heavily affected by the low-quality candidate augmentations, as shown in Tabel 8. On the other hand, by adaptively selecting optimal augmentations, InfoTS significantly improve the forecasting performances even with low-quality candidate augmentations. 2) By comparing InfoTS to its variant w/o Fidelity and w/o Variety, we further empirically demonstrate that both variety and fidelity make contributions to the superior performances of InfoTS.

D.4 FULL RESULTS OF TIME SERIES PREDICTION AND CLASSIFICATION

The full results of multivariate time series forecasting are shown in Tabel 4. Results of StemGNN with $L_y = 720$ are not available due to the out-of-memory error (Yue et al., 2021). Full results of univariate time series classification on 128 UCR datasets are shown in Tabel 9. Results of T-Loss, TS-TCC, and TNC are not reported on several datasets because they are not able to deal with missing observations in time series data. These unavailable accuracy scores are dismissed when computing average accuracy and considered as 0 when calculating the average rank. Results of multivariate

Table 8: Ablation studies with controllable augmentations.

L_y	InfoTS		Data Augmentation				Meta Objective			
			Random		All		w/o Fidelity		w/o Variety	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
24	0.246	0.269	0.252	0.284	0.249	0.279	0.250	0.274	0.251	0.274
48	0.294	0.298	0.303	0.316	0.303	0.312	0.301	0.308	0.297	0.303
168	0.400	0.369	0.414	0.386	0.414	0.388	0.404	0.369	0.409	0.376
336	0.542	0.455	0.565	0.477	0.563	0.480	0.554	0.456	0.542	0.353
Avg.	0.370	0.348	0.384	0.366	0.382	0.365	0.377	0.352	0.375	0.353

classification on 30 UEA datasets are listed in Tabel 10. Computations of average accuracy scores and ranks follow the ones in 128 UCR datasets.

	InfoTS _s	InfoTS	TS2Vec	T-Loss	TNC	TS-TCC	TST	DTW
Adiac	0.795	0.788	0.775	0.675	0.726	0.767	0.550	0.604
ArrowHead	0.874	0.874	0.857	0.766	0.703	0.737	0.771	0.703
Beef	0.900	0.833	0.767	0.667	0.733	0.600	0.500	0.633
BeetleFly	0.950	0.950	0.900	0.800	0.850	0.800	1.000	0.700
BirdChicken	0.850	0.900	0.800	0.850	0.750	0.650	0.650	0.750
Car	0.900	0.883	0.883	0.833	0.683	0.583	0.550	0.733
CBF	1.000	0.999	1.000	0.983	0.983	0.998	0.898	0.997
ChlorineConcentration	0.825	0.822	0.832	0.749	0.760	0.753	0.562	0.648
CinCEGTorso	0.896	0.928	0.827	0.713	0.669	0.671	0.508	0.651
Coffee	1.000	1.000	1.000	1.000	1.000	1.000	0.821	1.000
Computers	0.720	0.748	0.660	0.664	0.684	0.704	0.696	0.700
CricketX	0.780	0.774	0.805	0.713	0.623	0.731	0.385	0.754
CricketY	0.774	0.774	0.769	0.728	0.597	0.718	0.467	0.744
CricketZ	0.792	0.787	0.792	0.708	0.682	0.713	0.403	0.754
DiatomSizeReduction	0.997	0.997	0.987	0.984	0.993	0.977	0.561	0.967
DistalPhalanxOutlineCorrect	0.808	0.801	0.775	0.775	0.754	0.754	0.728	0.717
DistalPhalanxOutlineAgeGroup	0.763	0.763	0.727	0.727	0.741	0.755	0.741	0.770
DistalPhalanxTW	0.720	0.727	0.698	0.676	0.669	0.676	0.568	0.590
Earthquakes	0.821	0.821	0.748	0.748	0.748	0.748	0.748	0.719
ECG200	0.950	0.930	0.920	0.940	0.830	0.880	0.830	0.770
ECG5000	0.945	0.945	0.935	0.933	0.937	0.941	0.928	0.924
ECGFiveDays	1.000	1.000	1.000	1.000	0.999	0.878	0.763	0.768
ElectricDevices	0.691	0.702	0.721	0.707	0.700	0.686	0.676	0.602
FaceAll	0.929	0.929	0.805	0.786	0.766	0.813	0.504	0.808
FaceFour	0.864	0.818	0.932	0.920	0.659	0.773	0.511	0.830
FacesUCR	0.917	0.913	0.930	0.884	0.789	0.863	0.543	0.905
FiftyWords	0.809	0.793	0.774	0.732	0.653	0.653	0.525	0.690
Fish	0.949	0.937	0.937	0.891	0.817	0.817	0.720	0.920
FordA	0.925	0.915	0.948	0.928	0.902	0.930	0.568	0.555
FordB	0.795	0.785	0.807	0.793	0.733	0.815	0.507	0.620
GunPoint	1.000	1.000	0.987	0.980	0.967	0.993	0.827	0.907
Ham	0.848	0.838	0.724	0.724	0.752	0.743	0.524	0.467
HandOutlines	0.946	0.946	0.930	0.922	0.930	0.724	0.735	0.881
Haptics	0.545	0.546	0.536	0.490	0.474	0.396	0.357	0.377
Herring	0.703	0.656	0.641	0.594	0.594	0.594	0.594	0.531
InlineSkate	0.420	0.424	0.415	0.371	0.378	0.347	0.287	0.384
InsectWingbeatSound	0.664	0.639	0.630	0.597	0.549	0.415	0.266	0.355
ItalyPowerDemand	0.971	0.966	0.961	0.954	0.928	0.955	0.845	0.950
LargeKitchenAppliances	0.851	0.853	0.875	0.789	0.776	0.848	0.595	0.795
Lightning2	0.934	0.934	0.869	0.869	0.869	0.836	0.705	0.869
Lightning7	0.863	0.877	0.863	0.795	0.767	0.685	0.411	0.726
Mallat	0.967	0.974	0.915	0.951	0.871	0.922	0.713	0.934
Meat	0.967	0.967	0.967	0.950	0.917	0.883	0.900	0.933
MedicalImages	0.920	0.820	0.793	0.750	0.754	0.747	0.632	0.737
MiddlePhalanxOutlineCorrect	0.859	0.859	0.838	0.825	0.818	0.818	0.753	0.698
MiddlePhalanxOutlineAgeGroup	0.662	0.662	0.636	0.656	0.643	0.630	0.617	0.500
MiddlePhalanxTW	0.636	0.617	0.591	0.591	0.571	0.610	0.506	0.506
MoteStrain	0.854	0.871	0.863	0.851	0.825	0.843	0.768	0.835
NonInvasiveFetalECGThorax1	0.941	0.941	0.930	0.878	0.898	0.898	0.471	0.790
NonInvasiveFetalECGThorax2	0.943	0.944	0.940	0.919	0.912	0.913	0.832	0.865
OliveOil	0.933	0.933	0.900	0.867	0.833	0.800	0.800	0.833
OSULeaf	0.760	0.760	0.876	0.760	0.723	0.723	0.545	0.591
PhalangesOutlinesCorrect	0.826	0.826	0.823	0.784	0.787	0.804	0.773	0.728
Phoneme	0.272	0.281	0.312	0.276	0.180	0.242	0.139	0.228
Plane	1.000	1.000	1.000	0.990	1.000	1.000	0.933	1.000
ProximalPhalanxOutlineCorrect	0.924	0.927	0.900	0.859	0.866	0.873	0.770	0.784
ProximalPhalanxOutlineAgeGroup	0.883	0.883	0.844	0.844	0.854	0.839	0.854	0.805
ProximalPhalanxTW	0.849	0.844	0.824	0.771	0.810	0.800	0.780	0.761
RefrigerationDevices	0.624	0.624	0.589	0.515	0.565	0.563	0.483	0.464
ScreenType	0.510	0.493	0.411	0.416	0.509	0.419	0.419	0.397
ShapeletSim	0.856	0.856	1.000	0.672	0.589	0.683	0.489	0.650
ShapesAll	0.855	0.852	0.905	0.848	0.788	0.773	0.733	0.768
SmallKitchenAppliances	0.773	0.773	0.733	0.677	0.725	0.691	0.592	0.643
SonyAIBORobotSurface1	0.921	0.927	0.903	0.902	0.804	0.899	0.724	0.725
SonyAIBORobotSurface2	0.953	0.953	0.890	0.889	0.834	0.907	0.745	0.831
StarLightCurves	0.973	0.973	0.971	0.964	0.968	0.967	0.949	0.907
Strawberry	0.978	0.978	0.965	0.954	0.951	0.965	0.916	0.941
SwedishLeaf	0.954	0.950	0.942	0.914	0.880	0.923	0.738	0.792
Symbols	0.979	0.979	0.976	0.963	0.885	0.916	0.786	0.950
SyntheticControl	1.000	1.000	0.997	0.987	1.000	0.990	0.490	0.993
ToeSegmentation1	0.930	0.934	0.947	0.939	0.864	0.930	0.807	0.772
ToeSegmentation2	0.923	0.915	0.915	0.900	0.831	0.877	0.615	0.838
Trace	1.000	1.000	1.000	0.990	1.000	1.000	1.000	1.000
TwoLeadECG	0.999	0.998	0.987	0.999	0.993	0.976	0.871	0.905
TwoPatterns	1.000	1.000	1.000	0.999	1.000	0.999	0.466	1.000
UWaveGestureLibraryX	0.820	0.819	0.810	0.785	0.781	0.733	0.569	0.728
UWaveGestureLibraryY	0.745	0.736	0.729	0.710	0.697	0.641	0.348	0.634
UWaveGestureLibraryZ	0.768	0.768	0.770	0.757	0.721	0.690	0.655	0.658
UWaveGestureLibraryAll	0.966	0.967	0.934	0.896	0.903	0.692	0.475	0.892

	InfoTS _s	InfoTS	TS2Vec	T-Loss	TNC	TS-TCC	TST	DTW
Wafer	0.999	0.998	0.998	0.992	0.994	0.994	0.991	0.980
Wine	0.963	0.963	0.889	0.815	0.759	0.778	0.500	0.574
WordSynonyms	0.715	0.704	0.704	0.691	0.630	0.531	0.422	0.649
Worms	0.766	0.753	0.701	0.727	0.623	0.753	0.455	0.584
WormsTwoClass	0.818	0.857	0.805	0.792	0.727	0.753	0.584	0.623
Yoga	0.937	0.869	0.887	0.837	0.812	0.791	0.830	0.837
ACSF1	0.850	0.850	0.910	0.900	0.730	0.730	0.760	0.640
AllGestureWiimoteX	0.560	0.630	0.777	0.763	0.703	0.697	0.259	0.716
AllGestureWiimoteY	0.623	0.686	0.793	0.726	0.699	0.741	0.423	0.729
AllGestureWiimoteZ	0.633	0.629	0.770	0.723	0.646	0.689	0.447	0.643
BME	1.000	1.000	0.993	0.993	0.973	0.933	0.760	0.900
Chinatown	0.985	0.988	0.968	0.951	0.977	0.983	0.936	0.957
Crop	0.766	0.766	0.756	0.722	0.738	0.742	0.710	0.665
EOGHorizontalSignal	0.577	0.572	0.544	0.605	0.442	0.401	0.373	0.503
EOGVerticalSignal	0.459	0.459	0.503	0.434	0.392	0.376	0.298	0.448
EthanolLevel	0.710	0.712	0.484	0.382	0.424	0.486	0.260	0.276
FreezerRegularTrain	0.998	0.996	0.986	0.956	0.991	0.989	0.922	0.899
FreezerSmallTrain	0.991	0.988	0.894	0.933	0.982	0.979	0.920	0.753
Fungi	0.866	0.946	0.962	1.000	0.527	0.753	0.366	0.839
GestureMidAirD1	0.592	0.592	0.631	0.608	0.431	0.369	0.208	0.569
GestureMidAirD2	0.459	0.492	0.515	0.546	0.362	0.254	0.138	0.608
GestureMidAirD3	0.323	0.315	0.346	0.285	0.292	0.177	0.154	0.323
GesturePebbleZ1	0.895	0.802	0.930	0.919	0.378	0.395	0.500	0.791
GesturePebbleZ2	0.905	0.842	0.873	0.899	0.316	0.430	0.380	0.671
GunPointAgeSpan	0.997	1.000	0.994	0.994	0.984	0.994	0.991	0.918
GunPointMaleVersusFemale	1.000	1.000	1.000	0.997	0.994	0.997	1.000	0.997
GunPointOldVersusYoung	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.838
HouseTwenty	0.941	0.924	0.941	0.933	0.782	0.790	0.815	0.924
InsectEPGRegularTrain	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.872
InsectEPGSmallTrain	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.735
MelbournePedestrian	0.964	0.962	0.959	0.944	0.942	0.949	0.741	0.791
MixedShapesRegularTrain	0.940	0.935	0.922	0.905	0.911	0.855	0.879	0.842
MixedShapesSmallTrain	0.892	0.887	0.881	0.860	0.813	0.735	0.828	0.780
PickupGestureWiimoteZ	0.820	0.820	0.820	0.740	0.620	0.600	0.240	0.660
PigAirwayPressure	0.433	0.432	0.683	0.510	0.413	0.380	0.120	0.106
PigArtPressure	0.820	0.830	0.966	0.928	0.808	0.524	0.774	0.245
PigCVP	0.654	0.653	0.870	0.788	0.649	0.615	0.596	0.154
PLAID	0.356	0.355	0.561	0.555	0.495	0.445	0.419	0.840
PowerCons	0.995	1.000	0.972	0.900	0.933	0.961	0.911	0.878
Rock	0.760	0.760	0.700	0.580	0.580	0.600	0.680	0.600
SemgHandGenderCh2	0.939	0.944	0.963	0.890	0.882	0.837	0.725	0.802
SemgHandMovementCh2	0.833	0.836	0.893	0.789	0.593	0.613	0.420	0.584
SemgHandSubjectCh2	0.945	0.924	0.951	0.853	0.771	0.753	0.484	0.727
ShakeGestureWiimoteZ	0.920	0.920	0.940	0.920	0.820	0.860	0.760	0.860
SmoothSubspace	1.000	1.000	0.993	0.960	0.913	0.953	0.827	0.827
UMD	1.000	1.000	1.000	0.993	0.993	0.986	0.910	0.993
DodgerLoopDay	0.675	0.675	0.562	-	-	-	0.200	0.500
DodgerLoopGame	0.971	0.942	0.841	-	-	-	0.696	0.877
DodgerLoopWeekend	0.986	0.986	0.964	-	-	-	0.732	0.949
AVG	0.841	0.839	0.836	0.806	0.761	0.757	0.639	0.729
Rank	1.773	1.969	2.328	3.648	4.515	4.391	6.125	5.133

Table 9: Full results of univariate time series classification on 128 UCR datasets.

Dataset	InfoTS _s	InfoTS	TS2Vec	T-Loss	TNC	TS-TCC	TST	DTW
ArticulatoryWordRecognition	0.993	0.987	0.987	0.943	0.973	0.953	0.977	0.987
AtrialFibrillation	0.267	0.200	0.200	0.133	0.133	0.267	0.067	0.200
BasicMotions	1.000	0.975	0.975	1.000	0.975	1.000	0.975	0.975
CharacterTrajectories	0.987	0.974	0.995	0.993	0.967	0.985	0.975	0.989
Cricket	1.000	0.986	0.972	0.972	0.958	0.917	1.000	1.000
DuckDuckGeese	0.600	0.540	0.680	0.650	0.460	0.380	0.620	0.600
EigenWorms	0.748	0.733	0.847	0.840	0.840	0.779	0.748	0.618
Epilepsy	0.993	0.971	0.964	0.971	0.957	0.957	0.949	0.964
ERing	0.911	0.911	0.874	0.133	0.852	0.904	0.874	0.133
EthanolConcentration	0.323	0.281	0.308	0.205	0.297	0.285	0.262	0.323
FaceDetection	0.525	0.534	0.501	0.513	0.536	0.544	0.534	0.529
FingerMovements	0.620	0.580	0.480	0.580	0.470	0.460	0.560	0.530
HandMovementDirection	0.514	0.392	0.338	0.351	0.324	0.243	0.243	0.231
Handwriting	0.554	0.452	0.515	0.451	0.249	0.498	0.225	0.286
Heartbeat	0.771	0.722	0.683	0.741	0.746	0.751	0.746	0.717
JapaneseVowels	0.986	0.984	0.984	0.989	0.978	0.930	0.978	0.949
Libras	0.889	0.883	0.867	0.883	0.817	0.822	0.656	0.870
LSST	0.593	0.591	0.537	0.509	0.595	0.474	0.408	0.551
MotorImagery	0.610	0.630	0.510	0.580	0.500	0.610	0.500	0.500
NATOPS	0.939	0.933	0.928	0.917	0.911	0.822	0.850	0.883
PEMS-SF	0.757	0.751	0.682	0.676	0.699	0.734	0.740	0.711
PenDigits	0.989	0.990	0.989	0.981	0.979	0.974	0.560	0.977
PhonemeSpectra	0.233	0.249	0.233	0.222	0.207	0.252	0.085	0.151
RacketSports	0.796	0.855	0.855	0.855	0.776	0.816	0.809	0.803
SelfRegulationSCP1	0.887	0.874	0.812	0.843	0.799	0.823	0.754	0.775
SelfRegulationSCP2	0.572	0.578	0.578	0.539	0.550	0.533	0.550	0.539
SpokenArabicDigits	0.927	0.947	0.988	0.905	0.934	0.970	0.923	0.963
StandWalkJump	0.467	0.467	0.467	0.333	0.400	0.333	0.267	0.200
UWaveGestureLibrary	0.884	0.884	0.906	0.875	0.759	0.753	0.575	0.903
InsectWingbeat	0.465	0.465	0.466	0.156	0.469	0.264	0.105	-
Avg. ACC	0.724	0.711	0.704	0.658	0.670	0.668	0.617	0.629
Avg. Rank	2.133	2.633	2.933	3.733	4.233	4.033	4.867	4.233

Table 10: Full results of multivariate time series classification on 30 UEA datasets.