SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2026, Vol. 69, Iss. 5, 152101:1–152101:18
https://doi.org/10.1007/s11432-025-4559-7

Trend virtual adversarial training for semi-supervised time series classification

Qingyi PAN[1], Liyuan WANG[2*], Jingyi ZHANG[3], Jun ZHU[2] & Ning CHEN[4*]

[1]*Department of Statistics and Data Science, Tsinghua University, Beijing 100084, China*
[2]*Institute for AI, Beijing Information Science and Technology National Research Center, Tsinghua-Bosch Joint Center for Machine Learning, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*
[3]*School of Mathematical Sciences, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[4]*High Performance Computing Center, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

Received 13 February 2025/Revised 5 June 2025/Accepted 13 August 2025/Published online 13 April 2026

Abstract Time series data analysis plays an important role in numerous application domains, including medical diagnosis, solar energy forecasting, and autonomous vehicle systems. A key characteristic of such data is the scarcity of labeled samples compared to the abundance of available unlabeled data, which has driven increasing attention toward semi-supervised learning approaches for time series analysis from both research and industrial communities. The widely-used virtual adversarial training (VAT) encourages model predictions that are invariant to small input perturbations for a smooth distribution with better generalization. Although VAT performs well in vision and language tasks, directly applying it to time series classification may corrupt key trend information, reducing its effectiveness for semi-supervised learning with unlabeled data. To address the above challenges, we propose trend virtual adversarial training (tVAT), which combines trend information extracted by Gaussian blurring and Lasso-inspired adversarial perturbations to effectively leverage unlabeled data for better generalization. We further theoretically demonstrate that the perturbed input can flexibly explore sample space without introducing spike-like anomalous patterns. Empirical evaluations show tVAT's consistent superior performance over competing baseline approaches in semi-supervised time series classification, achieving performance gains of up to 10.73%.

Keywords semi-supervised time series classification, virtual adversarial training, interpretable deep learning

Citation Pan Q Y, Wang L Y, Zhang J Y, et al. Trend virtual adversarial training for semi-supervised time series classification. Sci China Inf Sci, 2026, 69(5): 152101, https://doi.org/10.1007/s11432-025-4559-7

1 Introduction

In recent years, time series analysis has garnered widespread interest in many critical areas, such as electrocardiogram data in medical diagnosis [1], accident driving scenarios in autonomous driving applications [2], reinforcement learning [3], and power calibration in solar energy [4]. These applications typically involve scarce labeled data and massive unlabeled data, making semi-supervised time series classification a promising research direction. Although some methods have been proposed [5–10], existing methods struggle to effectively leverage massive unlabeled data with scarce labeled data, resulting in overfitting where deep models fit the labeled data so well that they fail to generalize to unlabeled data.

To alleviate the overfitting issue, a commonly used strategy is called consistency regularization, which enforces that model predictions that should be invariant to small input perturbations to promote smoothness of the predictive distribution. In particular, virtual adversarial training (VAT) [11] has shown promising performance in some specific areas such as natural language processing [12]. VAT defines local distributional smoothness (LDS) within the predictive distribution to generate gradient-based adversarial perturbations. However, a straightforward extension of the VAT to semi-supervised time series classification has proven to be less effective [13]. This is because the implementation of original adversarial perturbations may introduce spike-like anomalous patterns to disrupt key trend patterns. Also, due to the fact that such adversarial perturbations are often task-specific, perturbations that work for images or text may not be appropriate for time-series data, leading to degraded performance over unlabeled data.

* Corresponding author (email: ly_wang94@126.com, ningchen@mail.tsinghua.edu.cn)

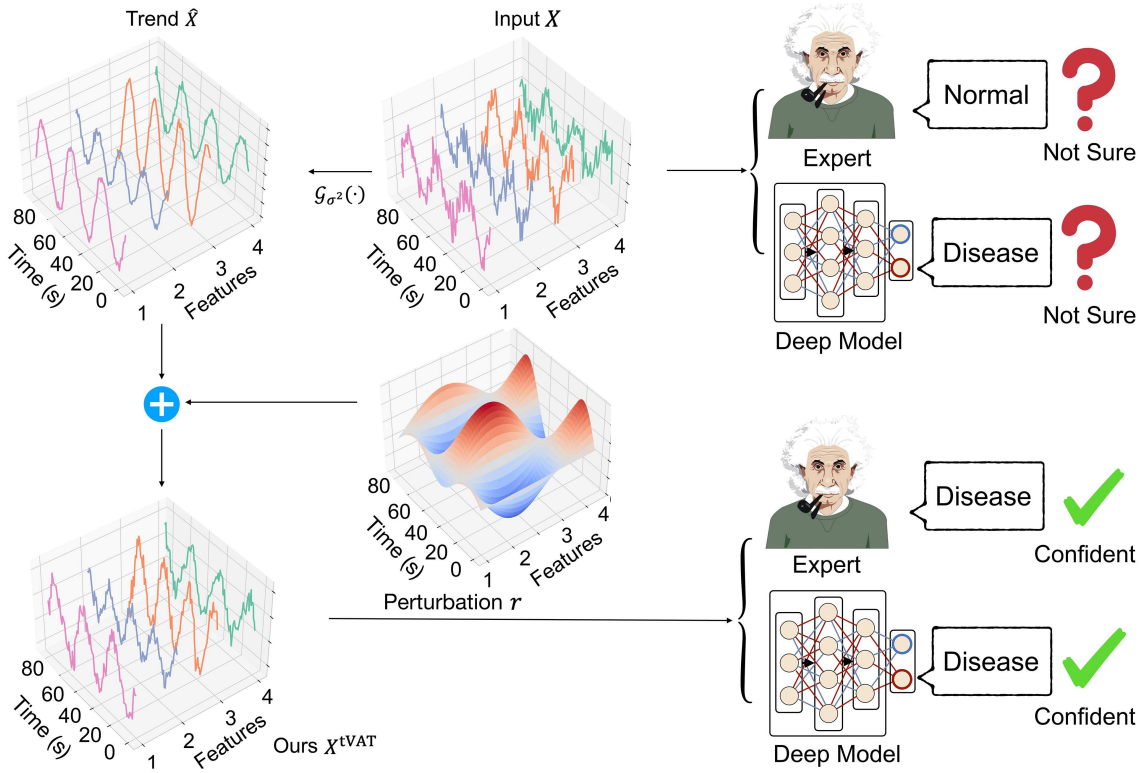

Figure 1 (Color online) The proposed tVAT effectively smooths spike-like anomalous patterns, while capturing key trend information in the heartbeat signal, thus helping medical experts in more accurate diagnoses.

To address the aforementioned challenges, we propose trend virtual adversarial training (tVAT) for semi-supervised time series classification. Specifically, we extract the trend information by Gaussian smoothing and introduce Lasso-inspired regularized adversarial perturbations to filter out spike-like anomalous patterns. We further theoretically demonstrate that the perturbed input flexibly explores the input space. These efforts enable deep models trained by tVAT to effectively use massive unlabeled data to yield a smooth predictive distribution with better generalization. We conduct extensive experiments across various semi-supervised time series classification tasks, including three univariate and three multivariate tasks. The empirical results verify that tVAT achieves superior performance compared to recent strong baselines. Furthermore, we provide qualitative and quantitative results to analyze the behaviors of deep models in semi-supervised settings. Figure 1 shows that the proposed tVAT attenuates rapid declines and abnormal peaks in heartbeat signals while preserving key periodic trend information, thus facilitating domain experts in analyzing behaviors of deep models. Furthermore, we conduct detailed ablation studies to verify the effectiveness of each component, such as an in-depth comparison of representative trend extraction methods.

In summary, our main contributions are as follows.

- We propose tVAT, which combines trend information extracted by Gaussian blurring with Lasso-inspired regularized adversarial perturbations to leverage unlabeled time series data for better generalization.
- We theoretically demonstrate that the Lasso-inspired regularized perturbed input flexibly explores the input space without introducing spike-like anomalous patterns.
- We conduct extensive experiments from qualitative and quantitative perspectives to verify the superiority of tVAT over recent baselines across various datasets.

The paper consists of the following sections. Section 2 reviews related work. Section 3 introduces the preliminaries of problem setups and VAT. Section 4 introduces the details of our proposed tVAT. Section 5 presents experimental results to verify the effectiveness of tVAT. Finally, Section 6 discusses limitations and future work.

2 Related work

In this section, we review related work in several key areas, including semi-supervised time series classification, trend analysis, and consistency regularization.

2.1 Semi-supervised time series classification

Semi-supervised time series classification appears in various applications such as diagnosis of medical diseases [14], electrocardiogram detection [1, 15], and reinforcement learning [3], which have received widespread interest from industry and academia in recent years. Traditional statistical methods use the nearest-neighbor algorithm with Euclidean distance for semi-supervised classification [16–18], yet are limited by their shallow architectures. In recent years, deep learning-based methods have been increasingly applied in these areas. Some classical deep semi-supervised learning methods from the image domain are transferred to the time series domain, but are limited by the significant domain gap [13]. MTL [6] adapts multi-task learning for semi-supervised settings, taking time series forecasting as an auxiliary task. Ref. [7] designs a novel architecture called TapNet, which integrates traditional and deep learning-based methods to improve feature representation. Recently, SemiTime [8] has proposed a self-supervised learning method, which matches past-future pairs to leverage unlabeled data. Class-aware temporal and contextual contrasting (CA-TCC) [10] applies weak/strong augmentations for unlabeled time series data and performs contrastive learning at the time/context level. Although these efforts reduce the dependence on labeled data, further exploration is needed to leverage large-scale unlabeled data for better generalization.

2.2 Trend analysis

Trend analysis in time series data has attracted extensive attention from both industry and academia [19–22]. The basic assumption is that each time series data can be decomposed into a trend component, a seasonal component, and noise [23]. In this regard, traditional methods, such as exponential moving average (EMA) [24], fast Fourier transform (FFT) [25], Savgol-filtering [26], and Gaussian smoothing [27], extract trend patterns from time series data. Similarly, deep models including recurrent neural networks [28] and convolutional neural networks [29] use trend information to improve performance in time series data. Recently, Semi-Time [8] incorporates trend information from unlabeled data to design additional self-supervised learning tasks, whose naive augmentation may introduce spike-like anomalous patterns to degenerate performance. Thus, further research in effectively using trend information to improve the performance of semi-supervised learning methods remains a critical and open challenge.

2.3 Consistency regularization

Both VAT and our proposed method fall into the category of consistency regularization [11], whose model predictions should be invariant to input perturbations, thus encouraging smoother predictive distributions with better generalization [11,12]. In this regard, adversarial training-based methods have recently been applied in supervised learning settings for specific time series datasets such as domestic financial markets [30,31]. While these efforts identify and exploit critical trend information for better robustness, we focus on semi-supervised time series classification by leveraging extensive unlabeled data for better generalization.

However, as shown in the Sections 3 and 4. Directly applying VAT to semi-supervised time series classification fails to effectively capture key trend information of time series data. These observations inspire us to develop an efficient method that prevents deep models from overfitting to limited labeled data while effectively utilizing trend information for better generalization.

3 Preliminaries

In this section, we introduce the necessary preliminaries for our proposed methods. Subsection 3.1 introduces the problem setup of semi-supervised time series classification. Subsection 3.2 introduces the direct extension of VAT.

3.1 Problem setup

Formally, we use $\{X_1, X_2, \ldots, X_N\} \in \mathcal{X}$ to denote the time series dataset, where $\mathcal{X}$ is the input space and $X_i \in \mathbb{R}^{T\times C}$ is the ith sample with length T and dimension C. The dataset is univariate if $C = 1$ and multivariate if $C > 1$. $X_{i,t}$ denotes the feature vector at the tth step of the ith feature vector, and each X_i has a categorical label $y_i \in \mathcal{Y} = \{1, 2, \ldots, K\}$. In semi-supervised time series classification, the partially labeled dataset $\mathcal{D} = \mathcal{D}^l \cup \mathcal{D}^u$ consists of a labeled subset $\mathcal{D}^l = \{(X_i^l, y_i)\}_{i=1}^{n_l}$ and an unlabeled subset $\mathcal{D}^u = \{X_i^u\}_{i=1}^{n_u}$, where $n_l \ll n_u$ and $n_l + n_u = n$. The goal is to learn a mapping function $f_\theta : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{D}$, with the conditional probability $p(y_i \mid X_i; \theta)$ parameterized by the learnable θ. The key trend information in time series data (i.e., the temporal dependencies of variables changing over time) often creates the complex input space and unsmooth predictive distribution of deep models, especially for non-stationary time series with inherent noise. Consequently, deep

models that are highly sensitive to small perturbations tend to overfit the labeled data and struggle to generalize to unlabeled or new data. Thus, there is an urgent need to improve the generalization of deep models by incorporating key trend information into adversarial perturbations.

3.2 VAT

In this subsection, we briefly review the original VAT [11], which improves the smoothness of predictive distributions by introducing perturbed input in local worst-case regions. Unlike traditional adversarial training limited to supervised settings, VAT effectively uses unlabeled data to generate adversarial perturbations. Specifically, the local distributional smoothness (LDS) is defined by the virtual adversarial loss:

$$\mathrm{LDS}(X_i, r_i; \theta) = D_{\mathrm{KL}}(p(\hat{y}_i \mid X_i; \theta) \mid\mid p(\hat{y}_i \mid X_i + r_i; \theta)), \tag{1}$$

where $p(\hat{y}_i \mid X_i, \theta)$ denotes the predictive distribution parameterized by learnable θ, and D_{KL} is the Kullback-Leibler (KL) divergence measuring the difference between two distributions [32]. The adversarial perturbation r_i has the same shape as the input X_i. We can approximate it by gradient descent iterations, since the closed form of r_i cannot be obtained directly,

$$r_i = \mathrm{argmax}_{\|r_i\|_2 \leqslant \epsilon} \mathrm{LDS}(X_i, r_i; \theta), \tag{2}$$

where ϵ is the hyperparameter controlling the perturbation norm determined by the validation set, and the average LDS loss over the entire dataset $\mathcal{D}$ is defined as

$$\mathcal{L}_{\mathrm{vadv}}(\mathcal{D}; \theta) = \frac{1}{|\mathcal{D}|} \sum_{X_i \in \mathcal{D}} \mathrm{LDS}(X_i, r_i; \theta). \tag{3}$$

On the other hand, the cross entropy loss function $\mathcal{L}_{\mathrm{CE}}$ on the labeled dataset $\mathcal{D}^l$ is defined as

$$\begin{aligned} \mathcal{L}_0(\mathcal{D}^l; \theta) &= \frac{1}{|\mathcal{D}|} \sum_{(X_i, y_i) \in \mathcal{D}^l} \mathcal{L}_{\mathrm{CE}}(X_i, y_i; \theta) \\ &= -\frac{1}{|\mathcal{D}|} \sum_{(X_i, y_i) \in \mathcal{D}^l} \sum_{j=1}^{K} y_{ij} \log p(\hat{y}_{ij} \mid X_i; \theta). \end{aligned} \tag{4}$$

The overall objective function consists of the supervised loss $\mathcal{L}_0$ on the labeled dataset $\mathcal{D}^l$ and the loss $\mathcal{L}_{\mathrm{vadv}}$ on the entire dataset $\mathcal{D}$ as

$$\mathcal{L} = \mathcal{L}_0\left(\mathcal{D}^l; \theta\right) + \mathcal{L}_{\mathrm{vadv}}\left(\mathcal{D}; \theta\right), \tag{5}$$

where the unique temporal dependencies of time series data limit the effectiveness of directly applying the original VAT. As r_i may introduce spike-like anomalous patterns that do not represent the worst-case perturbations, deep models trained with the original VAT may struggle to generalize on unlabeled data.

4 Our method

In this section, we introduce the proposed tVAT, including the detailed implementation in Subsection 4.1 and the fused Lasso regularizer with its theoretical properties in Subsection 4.2.

4.1 Trend VAT

To incorporate crucial trend information into the original VAT, we propose the tVAT, as shown in Figure 2. To facilitate further analysis, $\mathcal{L}_{\mathrm{vadv}}\left(\mathcal{D}; \theta\right)$ in (5) can be reformulated as

$$\mathcal{L}_{\mathrm{vadv}}(\mathcal{D}; \theta) = \frac{1}{|\mathcal{D}|} \sum_{X_i \in \mathcal{D}} \mathrm{LDS}\left(X_i, X_i^{\mathrm{tVAT}}; \theta\right), \tag{6}$$

where X_i^{tVAT} represents the perturbed input incorporating trend information for the ith sample X_i. We replace the KL divergence with the mean squared error, as a more effective logit learning strategy for LDS. Additional theoretical analysis in other applications, such as knowledge distillation [33] verifies that the mean squared error

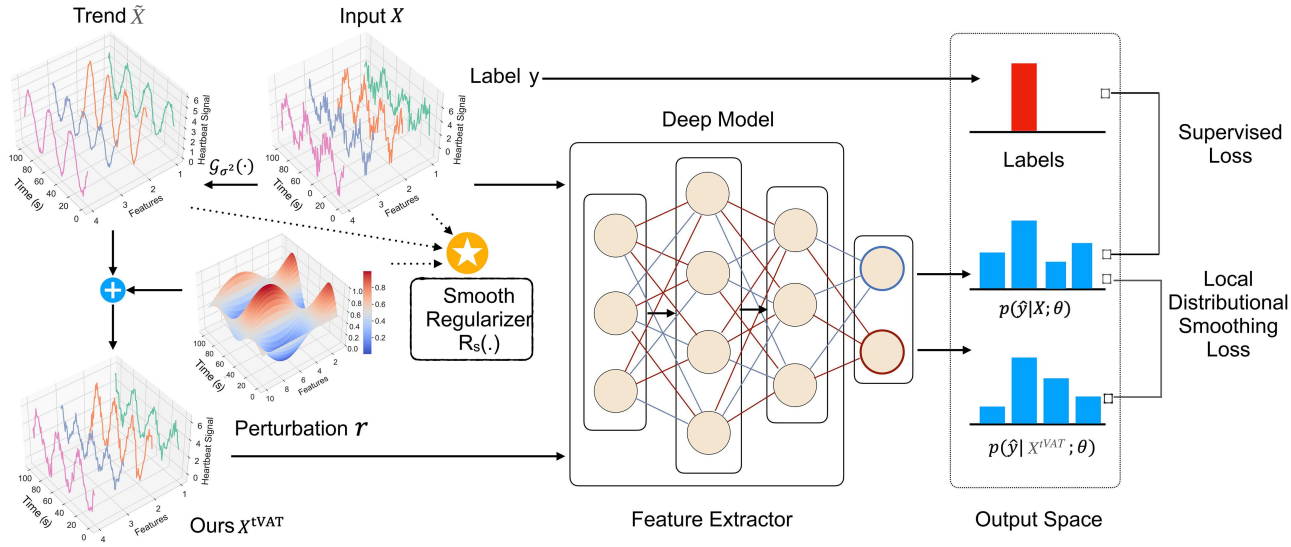

Figure 2 (Color online) The overall training procedure of tVAT combines trend information extracted by Gaussian smoothing with the Lasso-inspired regularized adversarial perturbation. Best viewed in color.

achieves better matching of logits or class probabilities than the KL divergence. The distance between X_i and X_i^{tVAT} is defined as

$$\mathrm{LDS}(X_i, X_i^{\mathrm{tVAT}}; \theta) = \|p(\hat{y}_i \mid X_i; \theta) - p(\hat{y}_i \mid X_i^{\mathrm{tVAT}}; \theta)\|_2^2. \tag{7}$$

To explicitly incorporate trend information into adversarial perturbation, the perturbed input X_i^{tVAT} combines the trend component $\tilde{X}_i$ with the regularized adversarial perturbation r_i^{tVAT} as

$$X_i^{\mathrm{tVAT}} = \tilde{X}_i + r_i^{\mathrm{tVAT}}, \tag{8}$$

where we extract the trend component $\tilde{X}_i$ by applying a Gaussian smoothing kernel $\mathcal{G}_{\sigma^2}(\cdot)$ [27] to the ith sample X_i [34]:

$$\tilde{X}_i = \mathcal{G}_{\sigma^2}(X_i), \tag{9}$$

where $\mathcal{G}_{\sigma^2}(\cdot)$ is the Gaussian smoothing kernel with isotropic noise variance σ^2 as a hyperparameter. The Gaussian smoothing kernel [27] filters out high-frequency anomalous patterns with only slight waveform distortion. $\mathcal{G}_{\sigma^2}(\cdot)$ balances trend information with detailed local information by setting a fixed optimal σ^2. The corresponding ablation in Subsection 5.4.2 also verifies the superiority of Gaussian smoothing over various representative trend extraction methods, including EMA [24], FFT [25], and Savgol filtering [26].

For the adversarial perturbation r_i^{tVAT}, we use the gradient approximation g_i to tackle the high-order optimization [35] as

$$\begin{aligned} r_i^{\mathrm{tVAT}} &= \eta \frac{g_i}{\|g_i\|_2}, \\ g_i &= \nabla_{r_i}\big[\mathrm{LDS}\big(X_i, X_i^{\mathrm{tVAT}}; \theta\big) + \mathcal{R}_s\big(X_i, X_i^{\mathrm{tVAT}}\big)\big], \end{aligned} \tag{10}$$

where η is the step size. The additional regularization term $\mathcal{R}_s(X_i, X_i^{\mathrm{tVAT}})$ adaptively smooths the perturbed input X_i^{tVAT} to avoid spike-like anomalous patterns and capture the underlying trend information:

$$\begin{aligned} \mathcal{R}_s(X_i, X_i^{\mathrm{tVAT}}) = \frac{1}{2}\sum_{t=1}^{T}\big(X_{i,t}^{\mathrm{tVAT}} - X_{i,t}\big)^2 \\ + \lambda \sum_{t=2}^{T-1}\big((X_{i,t-1}^{\mathrm{tVAT}} - X_{i,t}^{\mathrm{tVAT}}) + (X_{i,t+1}^{\mathrm{tVAT}} - X_{i,t}^{\mathrm{tVAT}})\big)^2, \end{aligned} \tag{11}$$

where T is the total number of time steps, and hyperparameter λ balances the two terms of $\mathcal{R}_s(X_i, X_i^{\mathrm{tVAT}})$. The first term measures the squared error distance between the ith input X_i and the perturbed input X_i^{tVAT}, while the second-order difference term evaluates the local smoothness of X_i^{tVAT}. The theoretical analysis in Subsection 4.2 ensures that the perturbed input X^{tVAT} effectively explores the sample space without introducing spike-like

Table 1 The definition of symbols in trend VAT.

Notation	Definition
X_i	The ith sample $X_i^{T\times C}$ with length T and feature dimension C
y_i	The class label for the ith sample
$\tilde{X}_i$	The trend information extracted by Gaussian smoothing $G_{\sigma 2}(X_i)$
X_i^{tVAT}	The perturbation integrating the trend information $\tilde{X}_i$ with r_i^{tVAT}
r_i^{tVAT}	The perturbation generated by (10)
$\mathcal{R}_s(\cdot)$	The Lasso-inspired regularization term
$f_\theta(\cdot)$	The deep model $f_\theta(\cdot)$ parameterized by learnable θ
λ	The hyperparameter to balance terms in the $\mathcal{R}(X_i, X_i^{\mathrm{tVAT}})$
η	The step size of stochastic gradient descent

anomalous patterns. The corresponding pseudo-code is provided in Algorithm 1. We also add a Nomenclature Table 1 to clarify key symbols of tVAT.

Furthermore, interpretability is another crucial aspect of semi-supervised settings, particularly for domain experts to analyze unlabeled time series data in some high-stakes areas like medical diagnosis [1] and self-driving applications [2]. The deep model trained by tVAT is expected to capture key features of time series data [36], aligning with previous work that deep models with better generalization learn meaningful feature representations closely aligned with human perception [37]. To further analyze the behaviors of deep models in semi-supervised time series classification, we raise some questions, such as "Can adversarial training more effectively extract key trend information?" or "Can adversarial training learn patterns across different label ratios?". The empirical results from both qualitative and quantitative perspectives in Subsection 5.3 show that tVAT effectively captures key trend information within time series data to further enhance generalization on the unlabeled dataset.

Algorithm 1 Trend VAT for semi-supervised time series classification.

Input: Dataset $\mathcal{D} = \mathcal{D}^l \cup \mathcal{D}^u$, hyperparameter λ, update steps N, number of epochs S, deep model f_θ parameterized by learnable θ.
Output: The optimized model parameters θ.
Initialize: Initialize model parameters θ.
while not converged **do**
  Sample $\{X_i^l, y_i\}$ from the labeled dataset $\mathcal{D}^l$, and calculate the cross-entropy loss in (4) over $\mathcal{D}^l$;
  Sample X_i from entire dataset $\mathcal{D}$;
  Calculate the regularizer $\mathcal{R}_s(X_i, X_i^{\mathrm{tVAT}})$ in (11);
  Calculate the $\mathrm{LDS}(X_i, X_i^{\mathrm{tVAT}}; \theta)$ in (7);
  Calculate the ith perturbation r_i^{tVAT} based on the $\mathrm{LDS}(X_i, X_i^{\mathrm{tVAT}})$ and $\mathcal{R}_s(X_i, X_i^{\mathrm{tVAT}})$ in (10);
  Calculate the trend pattern $\tilde{X}_i$ and perturbation X_i^{tVAT} for the ith sample X_i in (8) and (9);
  Calculate the unlabeled loss $\mathcal{L}_{\mathrm{vadv}}$ in (6) over the dataset $\mathcal{D}$;
  Calculate the total loss in (5) and update θ by stochastic gradient descent;
end while

4.2 Theoretical analysis

In this subsection, we analyze the Lasso-inspired regularization term $\mathcal{R}(X_i, X_i^{\mathrm{tVAT}})$ to provide a theoretical guarantee for the flexibility of virtual perturbations X^{tVAT}, suggesting that perturbation X^{tVAT} can effectively explore unexplored input spaces. To facilitate further analysis, the regularization term $\mathcal{R}(X_i, X_i^{\mathrm{tVAT}})$ for the ith sample X_i in (11) can be reformulated as

$$\frac{1}{2}\sum_{t=1}^{T}(X_{i,t} - X_{i,t}^{\mathrm{tVAT}})^2 + \lambda \sum_{t=2}^{T-1}(X_{i,t-1}^{\mathrm{tVAT}} - 2X_{i,t}^{\mathrm{tVAT}} + X_{i,t+1}^{\mathrm{tVAT}})^2, \tag{12}$$

where hyperparameter $\lambda \geqslant 0$ is crucial for balancing two terms in $\mathcal{R}_s(\cdot)$. The first term measures the discrepancy between X_i and X_i^{tVAT}, and the second-order difference term evaluates the local smoothness of the perturbed input. The objective function $\mathcal{J}(X_i^{\mathrm{tVAT}})$ for the ith perturbation X_i^{tVAT} is reformulated as a matrix form in (13). For simplicity, we denote $X = X_i$ and $X^{\mathrm{tVAT}} = X_i^{\mathrm{tVAT}}$ by omitting the subscript i for clarity.

$$\mathcal{J}(X^{\mathrm{tVAT}}) = \min_{X^{\mathrm{tVAT}}} \frac{1}{2}\|X - X^{\mathrm{tVAT}}\|_2^2 + \lambda \|MX^{\mathrm{tVAT}}\|_2^2, \tag{13}$$

where $\|u\|_2 = (\sum_t \mu_t^2)^{\frac{1}{2}}$ represents the l_2-norm, and $M \in \mathbb{R}^{(T-2)\times T}$ is the second-order difference operator as

$$M = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \ 1 \end{bmatrix}. \tag{14}$$

We set $\nabla \mathcal{J}(X^{\mathrm{tVAT}}) = 0$ with respect to X^{tVAT} as

$$X^{\mathrm{tVAT}} + 2\lambda M^\top M X^{\mathrm{tVAT}} = X. \tag{15}$$

Thus, the optimal condition for the perturbation X^{tVAT} is

$$X^{\mathrm{tVAT},*} = (I + 2\lambda M^\top M)^{-1} X. \tag{16}$$

Theorem 1. As $\lambda \to 0$, X^{tVAT} converges to the sample X with preserving diversity and satisfies the upper bound

$$\frac{\|X - X^{\mathrm{tVAT}}\|_2}{\|X\|_2} \leqslant \frac{32\lambda}{1 + 32\lambda}. \tag{17}$$

The proof details can be found in Appendix A.1.

Theorem 2. As $\lambda \to \infty$, X^{tVAT} converges to a straight line that best fits X as

$$X_t^{\mathrm{tVAT}} = \alpha + \beta t, \tag{18}$$

where the intercept α and slope β are in the least squared form,

$$\begin{cases} \alpha = \dfrac{S_2 S_X - S_1 S_{tX}}{T S_2 - S_1^2}, \\ \beta = \dfrac{T S_{tX} - S_1 S_X}{T S_2 - S_1^2}, \end{cases} \tag{19}$$

$$\begin{cases} S_1 = \displaystyle\sum_{t=1}^{T} t, \\ S_2 = \displaystyle\sum_{t=1}^{T} t^2, \\ S_X = \displaystyle\sum_{t=1}^{T} X_t, \\ S_{tX} = \displaystyle\sum_{t=1}^{T} t X_t. \end{cases} \tag{20}$$

The proof details can be found in Appendix A.2.

For small λ, the perturbation X^{tVAT} converges to the original data while preserving diversity under the theoretical upper bound. As λ increases, the perturbation X^{tVAT} becomes smoother, ultimately converging to a linear regression that best fits the samples in the least squares sense. The Lasso-inspired regularization term $\mathcal{R}_s(\cdot)$ provides a theoretical guarantee for the flexibility of virtual perturbations, verifying that X^{tVAT} incorporates key trend information without introducing anomalous spike-like patterns to efficiently explore sample spaces.

We further theoretically verify that the Lasso-inspired regularizer $\mathcal{R}_s(\cdot)$ helps gradient-based adversarial perturbations preserve low-frequency trend structure, thus stabilizing the training process in semi-supervised settings. To facilitate further analysis, the regularizer $\mathcal{R}_s(\cdot)$ in (11) can be reformulated as

$$\mathcal{R}_s\big(X_i, X_i^{\mathrm{tVAT}}\big) = \frac{1}{2}(d_i + r_i^{\mathrm{tVAT}})^\top A\,(d_i + r_i^{\mathrm{tVAT}}), \quad A = I + 2\lambda M \succ I, \tag{21}$$

where A is a positive-definite matrix. $d_i = \tilde{X}_i - X_i$ denotes the difference between the ith sample X_i and the trend information $\tilde{X}_i$.

Table 2 The statistics of univariate and multivariate datasets in the UEA & UCR archive, including three univariate datasets and three multivariate datasets.

Dataset	Samples	Length	Dim	Class
CricketX	780	300	1	12
UWaveGestureLibraryAll	4478	948	1	8
InsectWingbeatSound	2200	256	1	11
Heartbeat	409	405	61	5
NATOPS	360	51	24	6
SelfRegulationSCP2	380	1152	7	2

Theorem 3. Let $g_i = \nabla_{r_i}\big[\mathrm{LDS}(X_i, X_i^{\mathrm{tVAT}}; \theta) + \mathcal{R}_s(X_i, X_i^{\mathrm{tVAT}})\big]$ be the gradient in (10). Then the regularizer part can be expanded in the spectral domain as

$$\nabla_{r_i}\mathcal{R}_s = A\,(d_i + r_i^{\mathrm{tVAT}}) = \sum_{k=0}^{F-1}(1 + 2\lambda\mu_k)\,v_k, \tag{22}$$

where $\{v_k\}_{k=0}^{F-1}$ are the orthogonal basis satisfying $(\sum_k v_k = d_i + r_i^{\mathrm{tVAT}})$, and $\mu_k = 4\sin^4\big(\frac{\pi k}{T}\big)$ is the eigenvalue associated with the kth frequency band. The proof details can be found in Appendix A.3.

Because μ_k increases monotonically with respect to k, the factor $(1+2\lambda\mu_k)$ is larger in the high-frequency bands. Consequently, these components of the gradient g_i decay faster in the training process. Therefore, adversarial perturbations r of tVAT retain low-frequency trend structures, leading to a more stable training process with better generalization in semi-supervised settings.

5 Experimental results and discussion

In this section, we conduct comprehensive experiments to verify the effectiveness of our proposed tVAT. Subsection 5.1 describes the experimental setups, which include several real-world datasets. Subsection 5.2 compares tVAT with other competitive baselines in multiple real-world datasets in semi-supervised and fully supervised settings. Subsection 5.3 analyzes feature importance to interpret the behaviors of deep models from both quantitative and qualitative perspectives. Subsection 5.4 shows the ablation studies, including hyperparameter analysis and evaluation on various deep architectures to investigate the influence of each component in tVAT.

5.1 Experimental setup

5.1.1 *Datasets*

Following [8], we use six public univariate and multivariate time series datasets from the UCR & UEA time series archive [38] to evaluate the effectiveness of our proposed tVAT. These representative datasets, which range from simple to complex with dimensions from 1 to 61, are widely used in semi-supervised time series classification [8,13]. Each dataset is divided into training (60%), validation (20%), and test (20%) sets, and the raw datasets are scaled into the range [0,1] for stability. We report experimental results on univariate datasets (CricketX, UWaveGestureLibraryAll, InsectWingbeatSound) and more challenging multivariate datasets (Heartbeat, NATOPS, SelfRegulationSCP2). The statistics of these datasets are summarized in Table 2.

• **CricketX [39].** The dataset contains gesture data corresponding to the position of the X axis collected from accelerometers in 3D space. CricketX can be categorized into 12 classes, including gestures such as "Cancel Call", "Dead Ball", "Four", "Last Hour Leg Bye", "No Ball", "One Short", "Out", "Penalty Runs", "Six", "TV Replay". Both training and test sets contain 390 samples.

• **UWaveGestureLibraryAll [40].** The gesture data library contains 4478 samples collected by the Nokia search engine from users interacting with mobile phones through a gesture identification system. The training and test sets contain 896 and 3582 samples, respectively.

• **InsectWingbeatSound [41].** The Computational Entomology group at the University of California Riverside releases InsectWingbeatSound for insect classification, including male and female mosquitoes, two types of flies, and other insects. The training set contains 220 samples, and the test set contains 1980 samples. The overlap for each second is 70%, which refers to the portion of data window shared or reused across consecutive windows when segmenting time series data [23].

• **Heartbeat [42].** The Heartbeat dataset [42] released by the PhysioNet Challenge 2016 primarily includes heart sound signals from volunteers in clinical or non-clinical environments. The signals are categorized into two classes: "normal" (113 samples) and "abnormal" (296 samples). These sensors are placed in the aortic, pulmonic, tricuspid, and mitral areas of adults and children.

• **NATOPS [43].** The AALTD competition uses the NATOPS dataset [43]. These sensors collect data from the hands, elbows, wrists, and thumbs. The dataset consists of position coordinates. The six categories represent different actions: "I have to command", "all clear", "not clear", "spread wings", "fold wings", and "lock wings". Both training and test sets contain 180 samples.

• **SelfRegulationSCP2 [44].** The University of Tuebingen releases SelfRegulationSCP2 [44], which contains EEG data with seven columns and 1152 rows. These sensors record signals of slow cortical potentials from auditory and visual feedback. The training set contains 200 samples, and the test set contains 180 samples.

5.1.2 *Metrics*

Performance metric: accuracy. Following [8], we use accuracy as the primary metric to evaluate the performance of deep models in the semi-supervised setting. The metric in (23) measures the proportion of correct predictions relative to the total number of predictions, and is widely used in semi-supervised classification research [5–7],

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of all predictions}} \times 100\%. \tag{23}$$

Interpretability metric: AUSSL. To evaluate the interpretability of various semi-supervised learning methods, we propose a novel metric called the area under semi-supervised learning (AUSSL) to analyze the behaviors of deep models. Specifically, we generate feature importance maps $F(X)$ [45] for deep models and rank each feature according to its importance weight, where $F_e(X_{i,t})$ is the eth element in $\{F_e(X_{i,t})\}_{e=1}^{T\times C}$. The top $M\%$ features that satisfy $\frac{\sum_{e=1}^{s} F_e(X_{i,t})}{\sum_{i=1}^{C}\sum_{t=1}^{T} F_e(X_{i,t})} \approx M\%$ are perturbed by Gaussian noise $\mathcal{N}(\mu, \sigma^2)$, where $\mu = 0$ and $\sigma^2 = 0.1$. Then we adjust the percentage $[1\%, 2\%, \ldots, 10\%]$ to generate the performance degradation curve and calculate its area under the curve as AUSSL. A sharp decline in performance indicates that the perturbed features are critical for correct predictions, while minimal changes in performance indicate that the perturbed features have little impact on performance. Therefore, a lower AUSSL (i.e., the smaller area under the curve) indicates that the corresponding semi-supervised learning methods can effectively capture key features like trend information for predictions.

5.1.3 *Hyperparameters*

We use stochastic gradient descent with a learning rate of 10^{-2}. The batch size is set to 32 with a maximum of 300 epochs. We use Gaussian smoothing with isotropic variance $\sigma^2 = 0.1$ to extract trend information. Due to the model-agnostic properties of tVAT, we empirically compare various deep architectures in Subsection 5.4.3, and select an eight-layer temporal convolutional network (TCN) [46] as the primary backbone for evaluation. We run our experiments on eight NVIDIA A10 GPUs (each with 24 GB memory). We further perform a grid search to select tVAT-specific hyperparameters. We set $\sigma^2 = 0.1$ for Gaussian blurring from the hyperparameter list $[0.01, 0.1, 1, 5, 10]$ and set $\lambda = 1$ in (10) from the hyperparameter list $[0.001, 0.01, 0.1, 1, 10]$. More details can be found in the supplementary materials for reference.

5.2 Experimental results

5.2.1 *Semi-supervised performance*

Baseline. We compare tVAT with various strong baselines, including CA-TCC [10], SemiTime [8] and TapNet [7]. SupL only trains deep models on the labeled dataset. Pseudo-label [47] enlarges the labeled set with high-confidence pseudo labels. Π model [48] enforces prediction consistency across random augmentations (i.e., flip [34]) via self-ensembling. MTL [6] jointly optimizes a supervised classification task on the labeled dataset and an auxiliary self-supervised forecasting task on the unlabeled dataset. SemiTime [8] introduces a self-supervised task to predict temporal correlations and captures trend information to learn more discriminatory features. TapNet [7] refines the class prototypes of an attention-based network by leveraging embeddings from an unlabeled dataset to improve generalization. Class-Aware Temporal and Contextual Contrasting (CA-TCC) [10] applies weak and strong augmentations especially for time series data, and performs contrastive learning at time/context levels. The extra computational complexity of tVAT is $\mathcal{O}(N)$ with a small constant factor, consisting of extracting trend information

Table 3 The accuracy (%) of various methods on the univariate and multivariate datasets across label ratio $\alpha \in \{0.1, 0.2, 0.4\}$. We report the mean and standard deviation over five runs. The best performance is shown in boldface.

Dataset	Ratio	SupL	Pseudo-label [47]	Π model [48]	MTL [6]	SemiTime [8]	TapNet [7]	CA-TCC [10]	tVAT
CricketX	0.1	33.62±0.95	38.87±2.26	38.61±2.29	40.94±1.97	44.88±3.13	42.38±0.82	41.77±0.79	**47.94±3.57**
	0.2	38.79±2.08	44.44±2.91	48.18±2.07	50.12±1.22	51.61±0.66	51.79±0.51	52.62±0.97	**58.85±3.31**
	0.4	52.64±2.53	53.39±2.18	54.73±1.04	55.10±1.12	58.71±2.78	58.51±0.29	58.99±0.55	**69.72±0.65**
UWaveGesture	0.1	75.81±0.84	75.72±1.85	77.26±0.31	76.35±0.56	81.46±0.60	83.42±0.84	78.21±1.85	**92.85±0.49**
	0.2	81.53±0.54	81.66±0.74	82.87±0.64	81.77±0.94	84.57±0.49	86.92±0.76	83.07±0.39	**95.49±0.46**
	0.4	85.81±0.66	86.45±1.20	86.17±0.91	86.01±0.68	86.91±0.47	89.15±0.65	84.42±0.54	**97.17±0.17**
InsectWing	0.1	50.96±1.58	43.16±3.20	51.47±0.36	50.45±1.01	54.96±1.61	55.09±0.48	51.44±1.07	**61.44±1.75**
	0.2	55.95±0.76	48.35±1.81	56.14±1.32	56.43±0.88	59.01±1.56	60.38±0.86	58.01±0.93	**64.84±2.01**
	0.4	61.41±0.96	55.32±2.04	62.20±0.53	60.90±0.87	62.38±0.76	62.18±0.19	64.48±0.67	**66.51±1.71**
Heartbeat	0.1	65.99±0.68	70.27±0.70	64.07±0.50	71.79±0.61	70.31±2.64	68.19±1.52	73.69±1.19	**76.59±1.82**
	0.2	68.68±0.46	71.73±0.50	71.93±0.70	72.61±0.52	71.15±0.64	72.49±0.35	75.03±0.04	**75.79±0.68**
	0.4	72.31±0.53	73.28±0.70	73.73±0.16	74.95±0.24	72.04±0.59	74.57±0.50	76.42±0.47	**78.59±1.37**
NATOPS	0.1	63.40±0.90	70.90±3.40	73.90±0.40	73.85±0.74	75.48±3.43	73.60±1.38	72.22±1.30	**84.64±2.75**
	0.2	80.64±1.31	78.94±0.19	78.94±0.19	80.56±0.22	84.64±1.30	76.16±1.16	86.00±2.68	**91.41±0.78**
	0.4	82.75±1.36	90.67±1.51	87.17±1.99	81.29±0.31	92.99±0.76	81.96±1.80	88.86±2.15	**96.61±1.19**
SelfRegulation	0.1	43.40±1.29	44.93±0.19	45.29±0.55	50.25±0.56	50.81±1.16	50.67±0.55	54.76±1.29	**55.01±1.51**
	0.2	46.03±0.35	47.87±0.50	47.80±0.44	52.61±0.68	52.24±0.59	51.57±0.25	53.38±1.28	**56.53±1.05**
	0.4	49.33±0.65	52.37±0.26	52.31±0.32	55.41±0.29	53.63±0.97	53.28±0.65	57.09±1.66	**59.02±1.93**

via Gaussian smoothing in (9) and generating Lasso-inspired regularized adversarial perturbations in (10), so the extra computational costs remain the same order as the most recent baseline CA-TCC [10].

Following [8, 30], we randomly select samples with different proportions $\alpha \in \{0.1, 0.2, 0.4\}$ from the training set as labeled data. We report the mean and standard error of tVAT over 5 runs with different random seeds in Table 3. The best-performing models are selected by cross-validation in the validation set. As shown in Table 3, tVAT significantly outperforms other competitive baselines and SupL across all univariate and multivariate time series datasets. This is because adversarially perturbed input incorporating trend information facilitates deep models to leverage the unlabeled dataset for better performance. In contrast, the naive augmentation (e.g., local cropping or random shifts [34]) in other competitive self-supervised-based methods like the CA-TCC [10], SemiTime [8], Π model [48] can introduce spike-like anomalous patterns to degenerate performance. Moreover, as the label ratio α increases from 10% to 40% in univariate time series datasets, the deep model trained by tVAT achieves better performance (e.g., up to 10.73% improvement on CricketX and 8.02% on UWaveGestureLibraryAll with label ratio $\alpha = 0.4$), while in multivariate time series datasets, lower label ratios lead to more significant performance improvements (e.g., up to 9.16% on NATOPS with $\alpha = 0.1$). These observations indicate that tVAT uses additional labeled data to capture critical trend information in relatively simple univariate time series datasets, while for multivariate time series datasets, tVAT enables deep models to capture trend information even with limited labeled data.

Real-world applications. To empirically verify the scalability of tVAT in real-world scenarios, we conduct experiments on large-scale China securities index (CSI) datasets (i.e., CSI 50 and 500 futures) for predicting the direction (upward or downward) of futures prices [49, 50]. The dataset collects over 4.2×10^4 records spanning from 2020 to 2022, and each time step includes bid/ask prices and corresponding trading volumes. Table 4 presents the performance of various semi-supervised learning methods across different label ratios and shows that tVAT consistently outperforms other competitive baselines in all settings, especially on more volatile CSI 500 futures. This is because adversarial perturbations incorporate trend information for scaling up to more challenging real-world applications.

5.2.2 *Fully-supervised performance*

The proposed tVAT can be easily extended to fully supervised settings. We compare tVAT with several competitive supervised learning methods. ED [51] serves as the classical one-nearest-neighbor classifier based on the Euclidean distance. TapNet [7] and ShapeNet [5] are two deep learning-based methods using manually designed features like shapelets. Additionally, we include two non-neural network classification methods called ROCKET [52] and HiveCOTE [53]. The results of the baselines in Table 5 are taken from previous papers [5, 7, 51].

Table 4 The accuracy (%) of various semi-supervised learning methods on two large-scale China Securities Index (CSI) futures datasets with different label ratios. The best performance is shown in boldface.

Futures	Ratio	SupL	Pseudo-label [47]	Π model [48]	MTL [6]	SemiTime [8]	TapNet [7]	CA-TCC [10]	tVAT
50	10%	40.55±0.88	51.48±1.65	52.84±3.10	53.94±0.13	54.11±1.45	54.53±0.45	55.38±1.50	**58.88±0.77**
	20%	43.69±2.84	53.87±0.74	55.99±1.20	54.97±0.61	56.69±1.01	55.79±0.13	57.82±1.02	**60.68±0.16**
	40%	54.77±0.55	54.96±0.30	57.34±2.50	57.19±0.57	57.34±0.20	58.55±0.11	58.69±0.22	**62.32±1.15**
500	10%	32.21±1.79	40.59±3.36	35.44±0.99	38.77±0.59	37.52±0.70	40.15±1.21	39.31±0.38	**43.03±1.02**
	20%	35.38±1.06	41.05±4.04	38.74±0.73	42.50±0.92	45.57±0.59	44.42±1.55	47.42±1.11	**50.96±1.43**
	40%	43.19±0.17	46.09±2.61	43.13±0.34	46.04±0.12	49.36±2.34	50.70±0.57	52.76±0.35	**57.65±0.55**

Table 5 The performance comparison between tVAT and other competitive baselines in fully-supervised settings. We report the mean and standard deviation over five runs. The best performance is shown in boldface.

Dataset	HIVE-COTE [53]	ROCKET [52]	ED [51]	TapNet [7]	ShapeNet [5]	tVAT
CricketX	74.73±0.51	76.79±0.16	62.85±0.12	65.32±0.37	68.93±0.43	**76.91±0.28**
UWaveGestureLibraryAll	92.48±0.24	93.24±0.35	87.62±0.31	88.74±0.58	91.73±0.29	**98.23±0.81**
InsectWingbeatSound	61.68±0.16	64.19±0.34	61.76±0.39	67.42±0.18	66.34±0.18	**70.43±0.16**
Heartbeat	72.73±0.12	71.16±0.13	63.41±0.14	71.94±0.96	75.83±0.49	**76.34±0.47**
NATOPS	82.06±0.23	88.18±0.51	86.71±0.28	93.98±0.35	88.12±0.18	**98.47±0.77**
SelfRegulationSCP2	51.68±0.32	51.18±0.14	48.42±0.23	55.68±0.46	58.15±0.31	**61.55±0.12**

As shown in Table 5, tVAT consistently outperforms other methods in all settings, especially for multivariate datasets (i.e., up to 4.49% improvement on NATOPS and 4.99% on UWaveGestureLibraryAll). Overall, tVAT consistently outperforms other competitive baselines in both semi-supervised and fully-supervised time series classification. These observations indicate that adversarial perturbations that incorporate trend information in tVAT help enhance the smoothness of predictive distributions for better generalization.

5.3 Experimental results on interpretability

In this subsection, we conduct experiments to analyze the behaviors of deep models trained with various semi-supervised learning methods from both qualitative and quantitative perspectives. Specifically, we use the saliency map [54] to identify key features contributing to predictions.

5.3.1 *Qualitative analysis*

We analyze the more challenging medical dataset Heartbeat, which has received significant attention from medical researchers. Figure 3 shows that the deep model initially focuses on basic periodic features such as feature 49 with a low label ratio $\alpha = 0.1$. As the label ratio increases to $\alpha = 0.4$, the deep model progressively captures aperiodic feature patterns like feature 10 besides periodic features. We further transform the data from the time domain to the frequency domain using the fast Fourier transform [55], in order to verify the periodicity of different temporal features. These observations suggest that deep models trained by tVAT begin with simple periodic features and gradually perform complex or ambiguous tasks as more labeled data becomes available, which provides valuable insights into analyzing semi-supervised time series classification. Moreover, our method increases confidence in the applications of artificial intelligence for medical diagnosis. tVAT enables non-expert researchers to conduct preliminary evaluations of heart rate anomalies, while domain experts confirm diagnoses of heart arrhythmia through in-depth analysis.

5.3.2 *Quantitative comparison*

In this subsection, we conduct a quantitative analysis to evaluate the interpretability of various semi-supervised learning methods under different label ratios.

Quantitative analysis of various methods. Figure 4(a) shows the AUSSL performance curve with the label ratio $\alpha = 0.4$. When key features are sequentially removed based on the saliency map, all methods show a steep drop in performance, with tVAT showing the smallest area under the performance curve. Similarly, Table 6 shows the AUSSL for various semi-supervised methods across multiple datasets (i.e., Heartbeat, NATOPS, and InsectWingbeatSound) with the same label ratio $\alpha = 0.4$, where tVAT outperforms other competitive baselines due to its ability to effectively extract key trend information from unlabeled data.

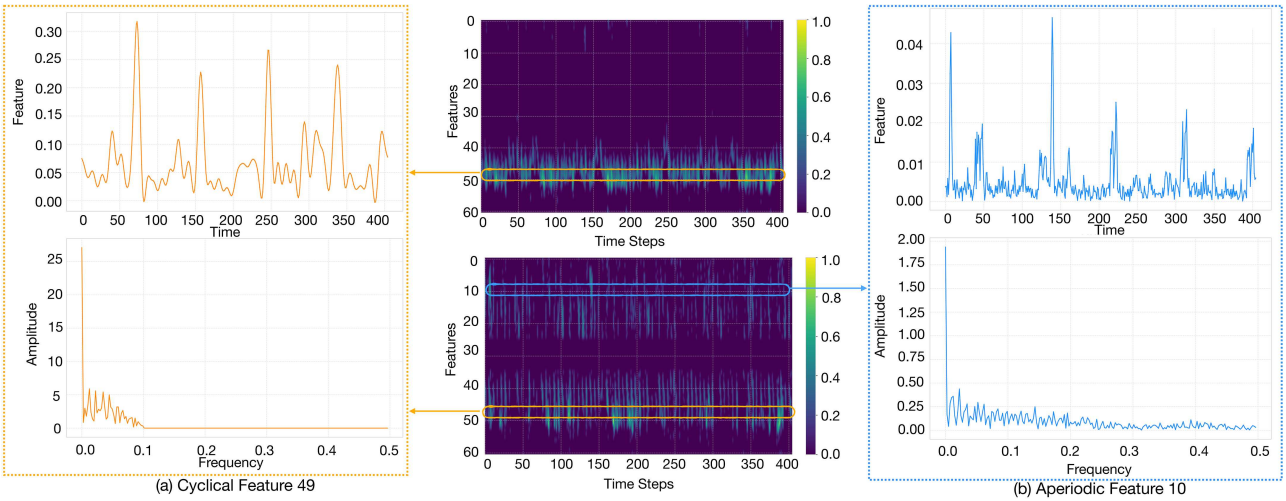

Figure 3 (Color online) The interpretable results on the Heartbeat dataset with label ratios $\alpha = 0.1$ and $\alpha = 0.4$. The deep model trained by tVAT initially learns the simpler cyclical feature 49, and progressively learns aperiodic feature 10 with $\alpha = 0.4$. These observations show that deep models begin with simple tasks and proceed into complex tasks as more labeled data becomes available. The results are best visualized in color.

Table 6 The AUSSL values of various baselines on InsectWingbeatSound, NATOPS, and Heartbeat datasets with a label ratio $\alpha = 0.4$. The best performance is shown in boldface.

Dataset	InsectWingbeatSound	NATOPS	Heartbeat
Pseudo-label [47]	0.0909	0.0316	0.0711
Π model [48]	0.0538	0.0326	0.0492
MTL [6]	0.0637	0.0338	0.0441
TapNet [7]	0.0796	0.0302	0.0424
Ours	**0.0312**	**0.0294**	**0.0325**

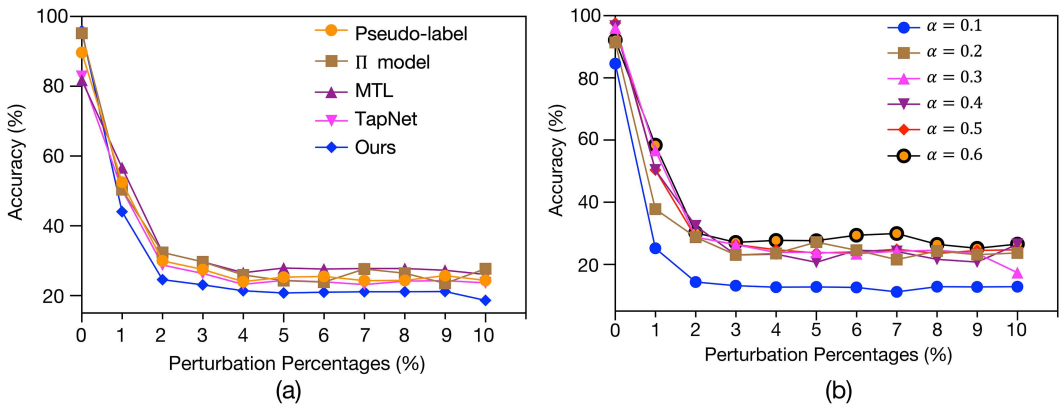

Figure 4 (Color online) (a) The accuracy curves of various semi-supervised learning methods with label ratio $\alpha = 0.4$ on the NATOPS dataset; (b) the accuracy curves of the deep model in tVAT with different label ratios $\alpha = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6]$ on the NATOPS dataset.

Quantitative analysis of label ratios. Figure 4(b) shows the performance curves of tVAT under different label ratios. For low label ratios like $\alpha = 0.1$, the performance curve decreases sharply, whereas high label ratios like $\alpha = 0.6$ stabilize the performance of deep models. These observations align with the qualitative analysis in Figure 3, where as the labeled ratio increases to provide stronger supervision signals, the deep model gradually focuses on capturing key features for predictions.

5.4 Ablation study

In this subsection, we conduct ablation studies to evaluate the effectiveness of each component in tVAT, including the hyperparameter analysis of λ in (11) and an investigation into the architectural choices of our method.

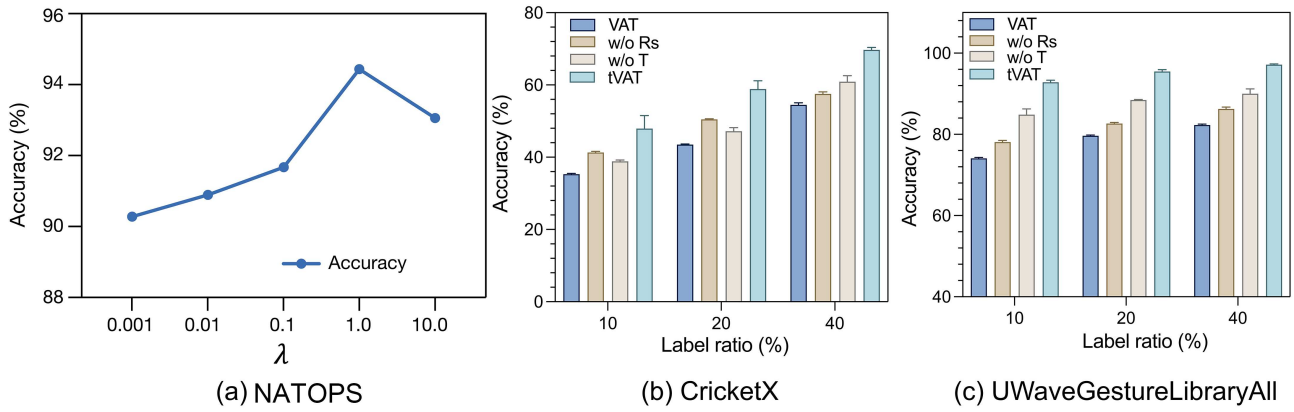

Figure 5 (Color online) (a) The performance curve by varying λ on NATOPS with the label ratio $\alpha = 0.4$; (b) the performance comparison of variants at different label ratios on the CricketX dataset; (c) the performance comparison of variants at different label ratios on the UWaveGestureLibraryAll dataset.

5.4.1 *Hyperparameter analysis*

For the design choices in (11), we conduct a parameter analysis of λ on the NATOPS with a label ratio $\alpha = 0.4$. We evaluate the performance of deep models trained by tVAT using different values of λ in the set $\{0.001, 0.01, 0.1, 10\}$. Figure 5(a) shows that tVAT achieves the best performance with $\lambda = 1$, indicating that both terms in $\mathcal{R}_s(\cdot)$ are equally important.

We construct additional ablation experiments to analyze contributions of trend extraction and regularization term $\mathcal{R}_s$. Specifically, we compare the performance of various variants in the ablation study. VAT presents the original VAT for semi-supervised time series classification. *w/o* $\mathcal{R}_s(\cdot)$ presents the deep model trained by tVAT without the regularization term $\mathcal{R}_s(\cdot)$ in (11). *w/o* T indicates that adversarial perturbations are applied directly to the samples instead of the corresponding trend information in (9), and tVAT denotes the complete version of our proposed method. As shown in Figures 5(b) and (c), we evaluate the performance of variants on several datasets (i.e., CricketX and UWaveGestureLibraryAll) across different label ratios. The empirical results show that both the regularization term $\mathcal{R}_s(\cdot)$ and the trend extraction module contribute to better performance, and removing each of them results in a drastic drop in performance.

5.4.2 *Trend extraction component*

We compare the performance across various trend extraction methods, including EMA [24], FFT [25], Savgol-filtering [26], and Gaussian smoothing [27]. Figure 6(a) shows that Gaussian smoothing consistently outperforms other trend extraction methods on CricketX across various label ratios, as it filters out high-frequency anomalous patterns with only slight waveform distortion and preserves the overall trend information for a more stable training process. Figure 6(b) presents the performance curve of Gaussian smoothing with different variance (i.e., $\sigma^2 \in \{0.01, 0.1, 1, 5, 10\}$). These observations show that the predictive performance peaks at $\sigma^2 = 0.1$ of Gaussian smoothing, because under-smoothing ($\sigma^2 < 0.1$) or over-smoothing ($\sigma^2 > 0.1$) both degenerate predictive performance. Based on this, we choose Gaussian smoothing with a specific empirical value $\sigma^2 = 0.1$ to extract trend information in the rest of experiments.

5.4.3 *Various architectures*

We compare different architectures trained by tVAT, including temporal convolutional network (TCN) [46], gated recurrent unit [56] and self-attention encoder (SA) [57]. These state-of-the-art architectures have shown their effectiveness in various sequence-modeling tasks.

• **Temporal Convolutional Network (TCN).** is a widely used architecture for sequence modeling, where its causal convolutional operations effectively capture both short-term fluctuations and long-term trend information [46]. By combining auto-regressive modules with long-term memory units, TCN outperforms recurrent architectures in many sequence modeling tasks.

• **Gated Recurrent Unit (GRU).** integrates the output and memory gates to address short-term memory challenges in sequence modeling [56]. GRU is often combined with attention mechanisms to effectively capture trend information.

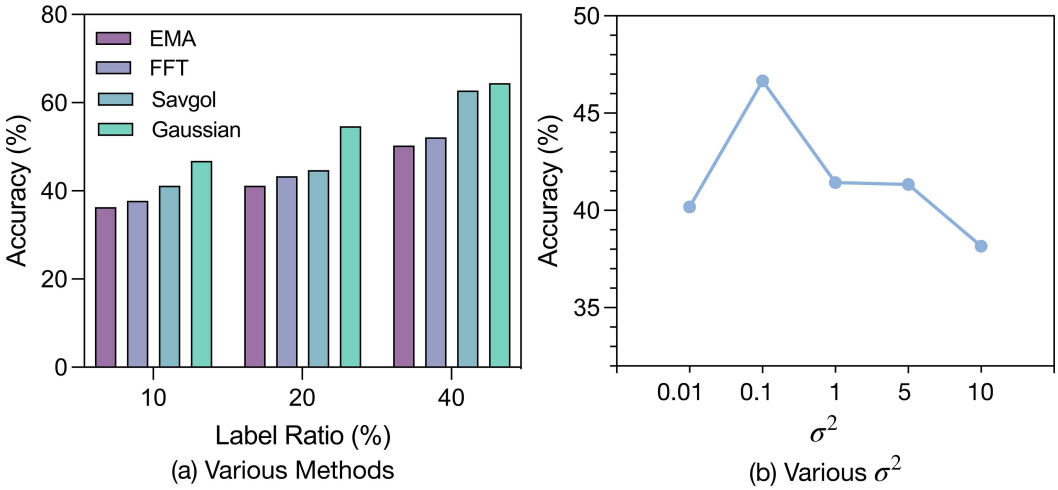

Figure 6 (Color online) (a) The performance comparison across various trend extraction methods; (b) the performance comparison across different σ^2 on CricketX.

Table 7 The accuracy (%) of tVAT based on various deep architectures with different label ratios. We report the mean and standard deviation over five runs. The best performance is shown in boldface.

Dataset	Ratio	GRU	SA	TCN
CricketX	10%	46.83±0.14	47.31±0.69	**48.32±0.46**
	20%	52.28±0.39	53.12±0.56	**54.17±0.30**
	40%	61.38±0.64	**63.94±0.18**	63.69±0.52
UWaveGestureLibraryAll	10%	86.33±0.24	88.06±0.96	**89.88±0.85**
	20%	88.68±0.13	90.19±0.98	**91.98±0.79**
	40%	90.84±0.73	92.14±0.13	**93.68±0.86**
InsectWingbeatSound	10%	54.12±0.14	**58.24±0.35**	57.91±0.97
	20%	60.72±0.25	61.12±0.45	**62.31±0.11**
	40%	62.57±0.28	**66.25±0.78**	65.97±0.52
Heartbeat	10%	69.91±0.22	71.22±0.97	**72.01±0.39**
	20%	71.62±0.71	73.35±0.68	**74.63±0.11**
	40%	74.23±0.66	76.24±0.41	**77.46±0.82**
NATOPS	10%	80.65±0.36	82.24±0.06	**83.93±0.77**
	20%	81.28±0.23	84.12±0.56	**85.61±0.91**
	40%	86.32±0.39	87.15±0.13	**89.58±0.78**
SelfRegulationSCP2	10%	53.01±0.55	51.13±0.33	**54.75±0.27**
	20%	54.43±0.12	53.44±0.87	**55.38±0.23**
	40%	53.12±0.22	55.12±0.62	**57.12±0.03**

• **Self-Attention encoder (SA) [57].** SA has recently been widely used in sequence modeling, such as natural language processing [58] and time series analysis [59]. The self-attention encoder, composed of four self-attention layers and positional encoding layers, effectively extracts long-term trend information from time series data.

Table 7 shows that TCN outperforms other deep architectures in most settings, except for the InsectWingbeatSound and CricketX datasets. The observations suggest that the causal convolutions effectively capture trend information, consistent with previous conclusions that effectively perceiving trend information enhances the smoothness of predictive distributions and improves generalization of deep models on unlabeled datasets. Additionally, Figure 7 shows the performance of TCN architectures with different numbers of layers (i.e., 8 or 10 layers) across multiple datasets at $\alpha = 0.4$, where the format of legend is "dataset (number of layers)". The empirical results show that the 8-layer TCN model provides sufficient capacity to capture key trend information.

6 Conclusion

In this paper, we introduce the tVAT for semi-supervised time series classification. tVAT incorporates trend information extracted by Gaussian smoothing into adversarial perturbations and introduces a specific Lasso-inspired

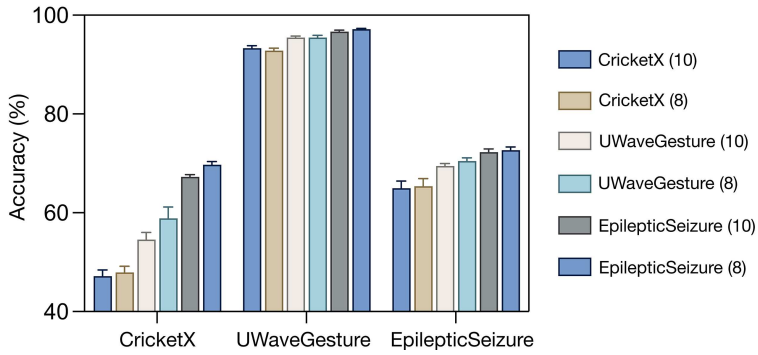

Figure 7 (Color online) The performance comparison of TCN architectures across multiple datasets with different layers.

regularization term to avoid anomalous patterns. We theoretically verify that the perturbed input can flexibly explore the input space without introducing spike-like anomalous patterns. The comprehensive empirical results show that our proposed tVAT performs remarkably better than other competitive baselines from both quantitative and qualitative perspectives. We conduct extensive ablation studies to verify the effectiveness of each component with our proposed tVAT. There are still some limitations in this study. We follow the implementations of representative semi-supervised time series classification settings, which primarily train all backbone models from scratch. Unfortunately, the limited computational resources prevent us from evaluating tVAT based on the most recent deep architectures like Mamba or TimeMixer. We consider examining the effectiveness of pre-trained or more powerful deep architectures as potential future work with sufficient computational resources.

Moreover, we observe a growing interest in the field of deep learning from the community of semi-supervised time series classification. Through an in-depth discussion of the potential challenges of VAT, we believe that this work can facilitate further exploration in this field and its intersection with consistency regularization-based methods.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62276149, 62406160, 92370124, 62350080, 92248303, U2341228, 62061136001), Beijing Information Science and Technology National Research Center (Grant No. BNR2022RC01-006) and Tsinghua Institute for Guo Qiang. Ning Chen is supported by the Tsinghua High Performance Computing Center and Tsinghua University-Bosch Joint Research Center for Machine-Learning. Jun Zhu is supported by the XPlorer Prize. Qingyi Pan and Jingyi Zhang are supported by National Natural Science Foundation of China (Grant No. 12301381), National Key R&D Program of China (Grant No. 2021YFA1001300), and Beijing Municipal Natural Science Foundation (Grant No. 1232019).

References

1 Zhai X, Zhou Z, Tin C. Semi-supervised learning for ECG classification without patient-specific labeled data. Expert Syst Appl, 2020, 158: 113411
2 Zhang J, Liu H, Lu J. A semi-supervised 3D object detection method for autonomous driving. Displays, 2022, 71: 102117
3 Qiaoben Y, Ying C Y, Zhou X N, et al. Understanding adversarial attacks on observations in deep reinforcement learning. Sci China Inf Sci, 2024, 67: 152104
4 Reikard G. Predicting solar radiation at high resolutions: a comparison of time series forecasts. Solar Energy, 2009, 83: 342–349
5 Li G, Choi B, Xu J, et al. ShapeNet: a shapelet-neural network approach for multivariate time series classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 8375–8383
6 Jawed S, Grabocka J, Schmidt-Thieme L. Self-supervised learning for semi-supervised time series classification. Adv Knowl Discov Data Mining, 2020, 12084: 499–511
7 Zhang X, Gao Y, Lin J, et al. TapNet: multivariate time series classification with attentional prototypical network. AAAI, 2020, 34: 6845–6852
8 Fan H, Zhang F, Wang R, et al. Semi-supervised time series classification by temporal relation prediction. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2021. 3545–3549
9 Cai R C, Wu Y J, Huang X K, et al. Granger causal representation learning for groups of time series. Sci China Inf Sci, 2024, 67: 152103
10 Eldele E, Ragab M, Chen Z, et al. Self-supervised contrastive representation learning for semi-supervised time-series classification. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 15604–15618
11 Miyato T, Maeda S I, Koyama M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans Pattern Anal Mach Intell, 2018, 41: 1979–1993
12 Lee D M, Kim Y, Seo C G. Context-based virtual adversarial training for text classification with noisy labels. ArXiv:2206.11851
13 Goschenhofer J, Hvingelby R, Rugamer D, et al. Deep semi-supervised learning for time series classification. ArXiv:2102.03622
14 Zhang Y, Li C, Liu Z, et al. Semi-supervised disease classification based on limited medical image data. IEEE J Biom Health Inform, 2024, 28: 1575–1586
15 Rasmussen S M, Jensen M E, Meyhoff C S, et al. Semi-supervised analysis of the electrocardiogram using deep generative models. In: Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, 2021. 1124–1127
16 Wei L, Keogh E. Semi-supervised time series classification. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, 2006. 748–753
17 Xu Z, Funaya K. Time series analysis with graph-based semi-supervised learning. In: Proceedings of the International Conference on Data Science and Advanced Analytics, 2015. 1–6
18 Chen Y, Hu B, Keogh E, et al. DTW-D: time series semi-supervised learning from a single example. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, 2013. 383–391
19 Ascari G, Sbordone A M. The macroeconomics of trend inflation. J Economic Literature, 2014, 52: 679–739
20 Eltsov T, Yutkin M, Patzek T W. Text analysis reveals major trends in exploration geophysics. Energies, 2020, 13: 4550
21 Pang Z, Berman O, Hu M. Up then down: bid-price trends in revenue management. Prod Oper Manag, 2015, 24: 1135–1147

22 Biondi F, Qeadan F. A theory-driven approach to tree-ring standardization: defining the biological trend from expected basal area increment. Tree-Ring Res, 2008, 64: 81–96
23 Hamilton J D. Time Series Analysis. Princeton: Princeton University Press, 2020
24 Randel W J. Filtering and data preprocessing for time series analysis. Statist Methods Phys Sci, 1994, 28: 283
25 Oppenheim A V. Discrete-Time Signal Processing. New York: Pearson Education India, 1999
26 Schafer R. What is a savitzky-golay filter? IEEE Signal Process Mag, 2011, 28: 111–117
27 Box G E P, Jenkins G M, Reinsel G C, et al. Time Series Analysis: Forecasting and Control. New York: John Wiley & Sons, 2015
28 Connor J T, Martin R D, Atlas L E. Recurrent neural networks and robust time series prediction. IEEE Trans Neural Netw, 1994, 5: 240–254
29 Koh B H D, Lim C L P, Rahimi H, et al. Deep temporal convolution network for time series classification. Sensors, 2021, 21: 603
30 Pialla G, Fawaz H I, Devanne M, et al. Smooth perturbations for time series adversarial attacks. In: Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2022. 485–496
31 Zhang Z, Li W, Bao R, et al. Asat: adaptively scaled adversarial training in time series. Neurocomputing, 2023, 522: 11–23
32 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
33 Kim T, Oh J, Kim N Y, et al. Comparing Kullback-Leibler divergence and mean squared error loss in knowledge distillation. International Joint Conference on Artificial Intelligence, 2021. 2628–2635
34 Wen Q, Sun L, Yang F, et al. Time series data augmentation for deep learning: a survey. ArXiv:2002.12478
35 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. ArXiv:1412.6572
36 Ismail A A, Gunady M, Bravo H C, et al. Benchmarking deep learning interpretability in time series predictions. ArXiv:2010.13924
37 Kim B, Seo J, Jeon T. Bridging adversarial robustness and gradient interpretability. ArXiv:1903.11626
38 Chen Y, Keogh E, Hu B, et al. The UCR time series classification archive. 2015. www.cs.ucr.edu/~eamonn/time_series_data/
39 Mueen A, Keogh E, Young N. Logical-shapelets: an expressive primitive for time series classification. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, 2011. 1154–1162
40 Liu J, Zhong L, Wickramasuriya J, et al. uWave: accelerometer-based personalized gesture recognition and its applications. Pervasive Mob Comput, 2009, 5: 657–675
41 Chen Y, Why A, Batista G, et al. Flying insect classification with inexpensive sensors. J Insect Behav, 2014, 27: 657–677
42 Liu C, Springer D, Li Q, et al. An open access database for the evaluation of heart sound algorithms. Physiol Meas, 2016, 37: 2181–2213
43 Ghouaiel N, Marteau P F, Dupont M. Continuous pattern detection and recognition in stream—a benchmark for online gesture recognition. Int J Appl Pattern Recognit, 2017, 4: 146–160
44 Birbaumer N, Ghanayim N, Hinterberger T, et al. A spelling device for the paralysed. Nature, 1999, 398: 297–298
45 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. ArXiv:1312.6034
46 Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. ArXiv:1803.01271
47 Lee D H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Proceedings of Workshop on Challenges in Representation Learning, 2013
48 Laine S, Aila T. Temporal ensembling for semi-supervised learning. ArXiv:1610.02242
49 Zhang Z, Zohren S, Roberts S. DeepLOB: deep convolutional neural networks for limit order books. IEEE Trans Signal Process, 2019, 67: 3001–3012
50 Pan Q, Sun S, Yang P, et al. FuturesNet: capturing patterns of price fluctuations in domestic futures trading. Electronics, 2024, 13: 4482
51 Bagnall A, Dau H A, Lines J, et al. The UEA multivariate time series classification archive. ArXiv:1811.00075
52 Dempster A, Petitjean F, Webb G I. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Min Knowl Discov, 2020, 34: 1454–1495
53 Bagnall A, Flynn M, Large J, et al. A tale of two toolkits, report the third: on the usage and performance of HIVE-COTE v1.0. In: Proceedings of Advanced Analytics and Learning on Temporal Data, 2020
54 Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the International Conference on Computer Vision, 2017. 618–626
55 Nussbaumer H J. The fast Fourier transform. In: Fast Fourier Transform and Convolution Algorithms. Berlin-Heidelberg: Springer, 1981, 80–111
56 Chung J, Gehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. ArXiv:1412.3555
57 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 5998–6008
58 Dai Z, Yang Z, Yang Y, et al. Transformer-XL: attentive language models beyond a fixed-length context. ArXiv:1901.02860
59 Li S, Jin X, Xuan Y, et al. Enhancing the locality and breaking the memory bottleneck of Transformer on time series forecasting. Adv Neural Inform Process Syst, 2019, 32: 5243–5253

Appendix A Proof

Appendix A.1 Proof of Theorem 1

Under the theoretical optimal condition for the perturbation in (16), we have

$$\begin{aligned} X - X^{\mathrm{tVAT}} &= 2\lambda M^\top M(I + 2\lambda M^\top M)^{-1}X \\ &= A(I + A)^{-1}X, \end{aligned} \tag{A1}$$

where $A = 2\lambda M^\top M$. Then, our goal is to bound $\|X - X^{\mathrm{tVAT}}\|_2 = \|A(I + A)^{-1}X\|_2$ using the sub-multiplicative property of norms[1)] in

$$\|X - X^{\mathrm{tVAT}}\|_2 \leqslant \|A(I + A)^{-1}\|_2 \|X\|_2. \tag{A2}$$

Then we need to calculate the eigenvalues of $\|A(I + A)^{-1}\|_2$. Since $M^\top M$ is symmetric and positive semi-definite, its eigenvalues $\{\lambda_i\}$ are non-negative and all satisfy $\{0 \leqslant \lambda_i \leqslant \lambda_{\max}\}$, so the eigenvalues of A and $(I + A)$ are $\mu_i = 2\lambda\lambda_i$ and $\nu_i = 1 + 2\lambda\lambda_i$. The eigenvalue of $A(I + A)^{-1}$ is

$$\eta_i = \frac{\mu_i}{\nu_i} = \frac{2\lambda\lambda_i}{1 + 2\lambda\lambda_i}, \tag{A3}$$

1) Strang G. Introduction to Linear Algebra. Philadelphia: SIAM, 2022.

where the operator norm of a symmetric matrix is equal to the absolute value of its largest eigenvalue[1)],

$$\|A(I+A)^{-1}\|_2 = \max_i \eta_i = \frac{2\lambda\lambda_{\max}}{1+2\lambda\lambda_{\max}}. \tag{A4}$$

Then we analyze the eigenvalues of $M^\top M$ for $\lambda_{\max}$. For the second-order difference operator M, its spectral norm $\|M\|_2$ is the square root of the largest eigenvalue of $M^\top M$,

$$\lambda_{\max} = \|M\|_2^2. \tag{A5}$$

For any sample $X \in \mathbb{R}^T$, we can use the sub-multiplicative property[1)] to calculate the norm of $\|MX\|_2$,

$$\begin{aligned} \|MX\|_2 &= \|X_{1:T-2} - 2X_{2:T-1} + X_{3:T}\|_2 \\ &\leqslant \|X_{1:T-2}\|_2 + 2\|X_{2:T-1}\|_2 + \|X_{3:T}\|_2 \\ &\leqslant 4\|X\|_2, \end{aligned} \tag{A6}$$

where the spectral norm $\|M\|_2 \leqslant 4$, so the largest eigenvalues of $M^\top M$ and the operator norm of $\|A(I+A)^{-1}\|_2$ is

$$\begin{aligned} &\lambda_{\max} = \|M\|_2^2 \leqslant 16, \\ &\|A(I+A)^{-1}\|_2 \leqslant \frac{2\lambda\times 16}{1+2\lambda\times 16} = \frac{32\lambda}{1+32\lambda}. \end{aligned} \tag{A7}$$

By substituting into (A2), the upper bound is

$$\frac{\|X - X^{\mathrm{tVAT}}\|_2}{\|X\|_2} \leqslant \frac{32\lambda}{1+32\lambda}. \tag{A8}$$

Appendix A.2 Proof of Theorem 2

As $\lambda \to \infty$, the regularizer $\lambda\|MX^{\mathrm{tVAT}}\|_2^2$ dominates the objective function in (13), and we only need to minimize $\|MX^{\mathrm{tVAT}}\|_2^2$. The term $\|MX^{\mathrm{tVAT}}\|_2^2$ is zero if and only if each $\{MX_t^{\mathrm{tVAT}}\} = 0$ for all t, which corresponds to homogeneous second-order difference equations as

$$X_{t-1}^{\mathrm{tVAT}} - 2X_t^{\mathrm{tVAT}} + X_{t+1}^{\mathrm{tVAT}} = 0, \quad t = \{2, \ldots, T-1\}, \tag{A9}$$

and the general solution to (A9) is

$$X_t^{\mathrm{tVAT}} = \alpha + \beta t. \tag{A10}$$

By substituting $X_t^{\mathrm{tVAT}} = \alpha + \beta t$, (A9) satisfies the optimal condition, i.e., $(\alpha+\beta(t-1)) - 2(\alpha+\beta t) + (\alpha+\beta(t+1)) = 0$. α and β can be derived by minimizing the ordinary least squares problem[2)],

$$\min_{\alpha,\beta} Q(\alpha,\beta) = \frac{1}{2}\sum_{t=1}^{T}(X_t - (\alpha+\beta t))^2, \tag{A11}$$

where the partial derivatives with respect to α and β is

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha} = -\displaystyle\sum_{t=1}^{T}(X_t - \alpha - \beta t) = 0, \\ \dfrac{\partial Q}{\partial \beta} = -\displaystyle\sum_{t=1}^{T} t(X_t - \alpha - \beta t) = 0, \end{cases} \tag{A12}$$

and the optimal conditions are formalized as a linear system in

$$\begin{cases} T\alpha + \beta\displaystyle\sum_{t=1}^{T} t = \sum_{t=1}^{T} X_t, \\ \alpha\displaystyle\sum_{t=1}^{T} t + \beta\sum_{t=1}^{T} t^2 = \sum_{t=1}^{T} tX_t, \end{cases} \tag{A13}$$

$$\begin{bmatrix} T & S_1 \\ S_1 & S_2 \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} S_X \\ S_{tX} \end{bmatrix}. \tag{A14}$$

Then we use Cramer's rule[1)] (i.e., determinant $\Delta \neq 0$) to solve α and β.

$$\Delta = TS_2 - (S_1)^2 = \frac{T^2(T+1)(T-1)}{12}, \tag{A15}$$

2) Casella G, Berger R. Statistical Inference. Boca Raton: CRC Press, 2024.

then α and β are derived by solving the linear equations in (A14),

$$\begin{cases} \alpha = \dfrac{S_2 S_X - S_1 S_{tX}}{T S_2 - S_1^2}, \\ \beta = \dfrac{T S_{tX} - S_1 S_X}{T S_2 - S_1^2}. \end{cases} \tag{A16}$$

As $\lambda \to \infty$, the perturbation X^{tVAT} converges to the best fit of the sample X in the least squares sense. The corresponding intercept α and slope β are derived by solving the linear system, facilitated by the strict convexity of the least squares objective function.

Appendix A.3 Proof of Theorem 3

For the ith time series of length T $X_i = (X_{i,1}, \ldots, X_{i,T})$ and its tVAT counterpart $X_i^{\mathrm{tVAT}} = (X_{i,1}^{\mathrm{tVAT}}, \ldots, X_{i,T}^{\mathrm{tVAT}})$, the trend regulariser is

$$R_s\big(X_i, X_i^{\mathrm{tVAT}}\big) = \frac{1}{2}\sum_{t=1}^{T}\big(X_{i,t}^{\mathrm{tVAT}} - X_{i,t}\big)^2 + \lambda \sum_{t=2}^{T-1}\big(X_{i,t-1}^{\mathrm{tVAT}} - 2X_{i,t}^{\mathrm{tVAT}} + X_{i,t+1}^{\mathrm{tVAT}}\big)^2, \tag{A17}$$

where the second term penalizes curvature (the discrete second derivative) and $\lambda > 0$ controls its strength. Then we reformulate the $(T-2) \times T$ second-difference matrix M as

$$M \;=\; \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix}, \tag{A18}$$

so that $\|M X_i^{\mathrm{tVAT}}\|_2^2 = \sum_{t=2}^{T-1}(X_{i,t-1}^{\mathrm{tVAT}} - 2X_{i,t}^{\mathrm{tVAT}} + X_{i,t+1}^{\mathrm{tVAT}})^2$. Then we introduce

$$r_i^{\mathrm{tVAT}} := X_i^{\mathrm{tVAT}} - X_i, \quad d_i := \tilde{X}_i - X_i, \quad z_i := d_i + r_i^{\mathrm{tVAT}},$$

and denote $B := M^\top M,\ A := I + 2\lambda B$. Because $B \succeq 0$ and $\lambda > 0$, matrix A is symmetric positive-definite. Substituting into (A17) yields the compact form

$$\mathcal{R}_s = \frac{1}{2}\, z_i^\top z_i + \lambda z_i^\top B z_i = \frac{1}{2}\, z_i^\top A z_i, \tag{A19}$$

which is strictly convex in z_i and r_i^{tVAT}. Differentiating (A19) with respect to the perturbation r_i^{tVAT} gives

$$\nabla_{r_i^{\mathrm{tVAT}}} \mathcal{R}_s\big(X_i, X_i^{\mathrm{tVAT}}\big) = A z_i = A\big(d_i + r_i^{\mathrm{tVAT}}\big). \tag{A20}$$

Thus, the complete gradient for tVAT is

$$g_i = \nabla_{r_i}\, \mathrm{LDS}\big(X_i, X_i^{\mathrm{tVAT}}; \theta\big) \;+\; A\big(d_i + r_i^{\mathrm{tVAT}}\big). \tag{A21}$$

For the symmetric Toeplitz matrix, we can take its eigenvectors as discrete cosine transform bases $\{v_k\}_{k=0}^{T-1}$ [55] with eigenvalues

$$\mu_k \;=\; 4\sin^4\left(\frac{\pi k}{T}\right), \quad k = 0, \ldots, T-1. \tag{A22}$$

Expanding $z_i = \sum_k \hat{z}_{i,k} v_k$ and applying (A20) we obtain

$$\nabla_{r_i^{\mathrm{tVAT}}} R_s = \sum_{k=0}^{T-1} (1 + 2\lambda\mu_k)\, \hat{z}_{i,k}\, v_k.$$

Hence, high-frequency components (k large) receive a larger multiplier $(1 + 2\lambda\mu_k)$, which suppresses rapid fluctuations while preserving low-frequency trend structure, leading to stable training and better generalization in semi-supervised learning settings.