
Functional Virtual Adversarial Training for Semi-Supervised Time Series Classification

Qingyi Pan

Department of Statistics and Data Science
Tsinghua University
Beijing, China
pqy22@mails.tsinghua.edu.cn

Yicheng Li*

Department of Statistics and Data Science
Tsinghua University
Beijing, China
liyc22@mails.tsinghua.edu.cn

Abstract

Real-world time series analysis, such as healthcare, autonomous driving, and solar energy, faces unique challenges arising from the scarcity of labeled data, highlighting the need for effective semi-supervised learning methods. While the Virtual Adversarial Training (VAT) method has shown promising performance in leveraging unlabeled data for smoother predictive distributions, straightforward extensions of VAT often fall short on time series tasks as they neglect the temporal structure of the data in the adversarial perturbation. In this paper, we propose the framework of functional Virtual Adversarial Training (f-VAT) that can incorporate the functional structure of the data into perturbations. By theoretically establishing a duality between the perturbation norm and the functional model sensitivity, we propose to use an appropriate Sobolev (H^{-s}) norm to generate structured functional adversarial perturbations for semi-supervised time series classification. Our proposed f-VAT method outperforms recent methods and achieves superior performance in extensive semi-supervised time series classification tasks (e.g., up to $\approx 9\%$ performance improvement). We also provide additional visualization studies to offer further insights into the superiority of f-VAT.

1 Introduction

Time series analysis has attracted considerable attention from both academia and industry, due to its relevance to critical domains such as electrocardiogram (ECG) interpretation in medical diagnosis [14] and photovoltaic module power calibration in solar energy systems [35]. However, capturing intrinsic temporal structural properties of time series data, such as noisy fluctuations, long-term trends, and periodic patterns, is still challenging [22], particularly with scarce labeled data [41]. Because manual annotation is labor-intensive and costly with massive unlabeled data, semi-supervised time series classification methods have become a promising research direction.

In this regard, existing methods attempt to leverage massive unlabeled data and limited labeled data to alleviate overfitting. Some studies generate high-confidence pseudo-labels for unlabeled samples and then train a deep model on the expanded dataset [6, 36]. Although such methods can iteratively improve predictive performance, they require careful selection of confidence thresholds and remain susceptible to “confirmation bias”: earlier errors in pseudo-labels are likely to be reinforced during training, making it difficult for deep models to correct early biases or capture global trend information [10, 2]. Some recent studies enhance pseudo-labeling with strong data augmentation and self-supervised representation learning to better utilize unlabeled data. For instance, TS-TCC [18] generates the pseudo-label on a weakly augmented view, while encouraging consistent representations

*Corresponding Author

under strong augmentation. However, naive data augmentation methods (e.g., local cropping, random shifting) [42] can disrupt key temporal information, thus affecting generalization.

Recently, Virtual Adversarial Training (VAT) based consistency regularization [31] constructs adversarial perturbations that maximally change the predictions of the model and then penalize the difference between the original and perturbed outputs, thus enforcing local smoothness and effectively tightening the decision boundary in the data manifold. While VAT and its variants have shown promising performance in the fields of computer vision [31], and natural language processing [28], most of these methods are based on the classical Euclidean norm to bound perturbations. Unfortunately, a straightforward extension of VAT to semi-supervised time series classification often proves less effective because bounding virtual perturbations via the Euclidean norm ignores the time series nature of the input data, producing jagged and spiky anomalous patterns that disrupt low-frequency trend information. Consequently, adversarial perturbations generated by standard VAT are not truly worst-case for time series data, making VAT inefficient in improving smoothness of predictive distributions and predictive performance.

To address these challenges, we propose *functional Virtual Adversarial Training (f-VAT)* for semi-supervised time series classification. F-VAT constructs adversarial perturbations in various (possibly) infinite-dimensional spaces, such as the Sobolev spaces [1], which can better capture the underlying structure of the data. For time series, this approach enables us to generate adversarial perturbations that can simultaneously preserve low-frequency trend information and flexibly explore input space, facilitating smoother predictive distribution and alleviating severe overfitting to limited labeled data.

In this paper, we theoretically establish the duality between the perturbation and the smoothness of (non-)linear functional models. Based on theoretical analysis, for time series data, we use an appropriate Sobolev norm to generate functional adversarial perturbations that preserve low-frequency trend information while avoiding “jagged” anomalous patterns. We conduct extensive experiments on real-world datasets to verify that f-VAT significantly outperforms other competitive baselines (e.g., up to 9.42% on CricketX and 8.30% on SelfReg). Further visualization indicates that, compared to the original VAT, our proposed functional adversarial perturbations lead to more stable convergence and better final performance.

In summary, our main contributions are as follows.

- We propose the framework of functional Virtual Adversarial Training (f-VAT) that allows us to construct perturbations in various function spaces. For linear and non-linear models, we theoretically establish the duality between the perturbation norm and gradient sensitivity (Theorem 3.3 and Theorem 3.4), showing how to generate structured adversarial perturbations.
- For the semi-supervised time series classification problem, we propose to use an appropriate Sobolev norm (H^{-s} norm in Section 3.3) to generate structured adversarial perturbations. The Sobolev norm allows us to properly control the model’s sensitivity to high-frequency noises while capturing low-frequency trend information.
- Extensive experimental results (Section 4.2) on semi-supervised time series classification are provided to demonstrate the superiority of f-VAT over existing methods. Additional visualization results (Section 4.3) indicate that functional adversarial perturbations can significantly smooth the loss landscape to achieve stable convergence and better performance.

2 Preliminaries on Virtual Adversarial Training

Let us briefly introduce the procedure of Virtual Adversarial Training (VAT) [31]. We consider the semi-supervised classification problem with a small labeled dataset and a large unlabeled dataset. Let $\{X_1, X_2, \dots, X_n\} \in \mathcal{X}$ denote the entire input data, where each X_i can be associated with a label $y_i \in \mathcal{Y} = \{1, 2, \dots, K\}$ if it is labeled, or y_i is unknown if it is unlabeled. We denote by $\mathcal{D}^l$ the labeled dataset, $\mathcal{D}^u$ the unlabeled dataset, and $\mathcal{D} = \mathcal{D}^l \cup \mathcal{D}^u$ the entire dataset. Our goal is to learn the conditional probability (vector) $p(\cdot | X_i; \theta) = f_\theta(X) \in \mathbb{R}^K$.

The idea of VAT is to generate adversarial perturbations r_i for each sample X_i and then penalize the difference between the original and perturbed outputs, facilitating the local smoothness of predictive distribution. Formally, the Local Distributional Smoothness $\text{LDS}(X_i, r_i; f_\theta)$ is used to measure such

difference for the model outputs. In this paper, we mainly focus on the squared loss

$$\text{LDS}(X_i, r_i; f_\theta) = \|f_\theta(X_i + r_i) - f_\theta(X_i)\|_2^2, \quad (1)$$

while another common choice is the Kullback-Leibler (KL) divergence [31]. We adopt the squared error, because its symmetry yields better numerical stability in the early stages of training [13]. The adversarial perturbation r_i aims to maximize the LDS loss, characterizing the worst-case local perturbation for the model f_θ as

$$r_i^* = \arg \max_{r_i: \|r_i\|_2 \leq \epsilon} \text{LDS}(X_i, r_i; f_\theta), \quad (2)$$

where $\epsilon > 0$ is a small hyperparameter controlling the perturbation norm in Euclidean space. Then, the VAT loss over X_i is defined as

$$\mathcal{L}^{\text{VA}}(X_i; f_\theta) = \text{LDS}(X_i, r_i^*; f_\theta) = \max_{r_i: \|r_i\|_2 \leq \epsilon} \text{LDS}(X_i, r_i; f_\theta). \quad (3)$$

In practice, since the analytic expression of r_i is intractable, an interactive gradient ascent [31] is used to approximate the optimal perturbation r_i^* . The overall objective function combines the supervised loss $\mathcal{L}_0$ on the labeled set $\mathcal{D}^l$ and the VAT loss $\mathcal{L}^{\text{VA}}$ on the entire dataset $\mathcal{D}$ is

$$\mathcal{L}(\mathcal{D}; f_\theta) = \mathcal{L}_0(\mathcal{D}^l; f_\theta) + \mathcal{L}^{\text{VA}}(\mathcal{D}; f_\theta), \quad (4)$$

where $\mathcal{L}^{\text{VA}}(\mathcal{D}; f_\theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}^{\text{VA}}(X_i; f_\theta)$ is the VAT loss over the entire dataset $\mathcal{D}$.

3 Functional Virtual Adversarial Training

One key aspect of VAT is the choice of the norm used to bound the local perturbation $\|r\| \leq \epsilon$ in Eq. (2). In the literature [31], it seems to be straightforward to use the Euclidean norm $\|\cdot\|_2$ as the input is typically represented vector in $\mathbb{R}^d$. However, when the input space $\mathcal{X}$ consists of functions, which is the case for time series data, the choice of the norm can become more intricate. As we will see in the following theoretical analysis, the choice of the norm can have a significant impact on the VAT loss and its interpretation. In this section, for simplicity, we consider one-dimensional output $f: \mathcal{X} \rightarrow \mathbb{R}$ and focus on some fixed sample x_0 , while our results can be easily extended to the case where f is a vector-valued function.

3.1 The Duality Perspective of VAT

To start with, let us first consider the setting where the input x is a vector in $\mathbb{R}^d$, but we consider a more general setting where the perturbation is bounded by a general norm. Let Σ be a positive definite matrix, we denote the Σ -norm by $\|v\|_\Sigma := \sqrt{v^\top \Sigma v}$. Replacing the Euclidean norm in Eq. (2) with the Σ -norm, we introduce the following VAT loss:

$$\mathcal{L}_\Sigma^{\text{VA}}(x_0, \epsilon; f) = \sup_{r: \|r\|_\Sigma \leq \epsilon} \text{LDS}(x_0, r; f). \quad (5)$$

To illustrate the impact of the Σ -norm on the VAT loss, let us first consider the case where $f(x)$ is a linear function. We have the following result.

Proposition 3.1. *Let $f(x) = \langle \beta, x \rangle$ be a linear model. Then, the VAT loss Eq. (5) is equivalent to*

$$\mathcal{L}_\Sigma^{\text{VA}}(x_0, \epsilon; f) = \epsilon^2 \|\beta\|_{\Sigma^{-1}}^2. \quad (6)$$

Proof. Since $f(x)$ is linear, we have

$$\mathcal{L}_\Sigma^{\text{VA}}(x_0, \epsilon; f) = \sup_{r: \|r\|_\Sigma \leq \epsilon} |\langle \beta, x_0 \rangle - \langle \beta, x_0 + r \rangle|^2 = \sup_{r: \|r\|_\Sigma \leq \epsilon} |\langle \beta, r \rangle|^2.$$

This supremum is achieved when r is in the direction of $\Sigma^{-1}\beta$, i.e.,

$$r^* = \epsilon \Sigma^{-1} \beta / \|\Sigma^{-1} \beta\|_\Sigma = \epsilon \Sigma^{-1} \beta / \|\beta\|_{\Sigma^{-1}},$$

and thus

$$\mathcal{L}_\Sigma^{\text{VA}}(x_0, \epsilon; f) = \sup_{r: \|r\|_\Sigma \leq \epsilon} |\langle \beta, r \rangle|^2 = \langle \beta, r^* \rangle^2 = \epsilon^2 \|\beta\|_{\Sigma^{-1}}^2.$$

□

Moreover, the result can be extended to the case where $f(x)$ is a continuously differentiable function as follows. The proof is deferred to Appendix A.3.

Proposition 3.2. *Let $f(x)$ be continuously differentiable. Then, we have*

$$\lim_{\epsilon \rightarrow 0^+} \epsilon^{-2} \mathcal{L}_{\Sigma}^{\text{VA}}(x_0, \epsilon; f) = \|\nabla f(x_0)\|_{\Sigma^{-1}}^2. \quad (7)$$

The results from Proposition 3.1 and Proposition 3.2 reveal a simple but profound duality structure underlying the VAT loss. First, as the gradient $\nabla f(x_0)$ reflects the sensitivity of the function, the VAT loss can be interpreted as a measure of the model's sensitivity to perturbations in the input space. Moreover, the use of the Σ -norm in the perturbation constraint $\|r\|_{\Sigma} \leq \epsilon$ naturally introduces a dual norm $\|\cdot\|_{\Sigma^{-1}}$ penalizing the sensitivity of f along the different directions. This is reminiscent of duality in optimization, where constraints in the primal space translate to penalties in the dual space.

Therefore, one of the crucial questions is how to choose the matrix Σ when designing the VAT loss. If Σ is isotropic (e.g., $\Sigma = I$), the VAT loss reduces to the Euclidean sensitivity $\|\nabla f(x_0)\|_2^2$. If Σ is anisotropic, the VAT loss prioritizes robustness along directions where Σ assigns lower curvature. This coincides with our intuition: allowing larger adversarial perturbations results in more constraint on the model's sensitivity in those directions. While it seems to be trivial when Σ is a diagonal matrix that simply scales the input, the choice of Σ can be more complex and encompasses structural information about the input space, which would lead to substantial differences in the VAT loss. This is what we will see under the time series data.

3.2 VAT under Functional Inputs

In the setting of the time series data [24], the input x is viewed as a function of some interval T , which would require a more sophisticated treatment on the theoretical analysis. Let $L^2 = L^2(T)$ be the Hilbert space of square integrable functions on T and denote the inner product by $\langle \cdot, \cdot \rangle_{L^2(T)}$.

We suppose that the input $x \in L^2(T)$. For the perturbation, let us take another Banach space E and denote by $\|\cdot\|_E$ the norm in E . Typically, we can choose E to be a Sobolev space, which will be discussed in the next subsection. The dual space E^* of E is defined as the space of continuous linear functionals on E , which is equipped with the dual norm $\|\cdot\|_{E^*}$ defined by

$$\|g\|_{E^*} = \sup_{x \in E: \|x\|_E \leq 1} |g(x)|.$$

To avoid technicalities, we will only provide informal results here. For a formal treatment and rigorous proofs, we refer the readers to Appendix A in appendix.

The functional virtual adversarial training (f-VAT) introduces a more general form of the VAT loss, which is defined as

$$\mathcal{L}_E^{\text{VA}}(x_0, \epsilon; f) = \sup_{r \in L^2 \cap E: \|r\|_E \leq \epsilon} \text{LDS}(x_0, r; f). \quad (8)$$

Our first result parallels Proposition 3.1 in the functional setting.

Theorem 3.3 (Informal). *Let $f(x) = \langle \beta, x \rangle_{L^2(T)}$, $\beta \in L^2(T)$ be a functional linear model. Then, the loss Eq. (8) is equivalent to*

$$\mathcal{L}_E^{\text{VA}}(x_0, \epsilon; f) = \epsilon^2 \|\beta\|_{E^*}^2, \quad (9)$$

where $\|\beta\|_{E^*}$ can be infinite if $\beta \notin E^*$.

Then, the next result is the functional version of Proposition 3.2.

Theorem 3.4 (Informal). *Let $f : L^2(T) \rightarrow \mathbb{R}$ be differentiable at $x_0 \in L^2(T)$. Then,*

$$\lim_{\epsilon \rightarrow 0^+} \epsilon^{-2} \mathcal{L}_E^{\text{VA}}(x_0, \epsilon; f) = \|\nabla f(x_0)\|_{E^*}^2. \quad (10)$$

Theorem 3.3 and Theorem 3.4 establish the duality of VAT under general functional spaces, showing that a constraint over the E -norm of the perturbation results in penalizing the dual space E^* -norm of the model's sensitivity. These results also recover Proposition 3.1 and Proposition 3.2 as special cases, since $\|\cdot\|_{\Sigma}$ and $\|\cdot\|_{\Sigma^{-1}}$ are dual under the canonical Euclidean norm. While finite dimensional spaces under different norms are equivalent, the infinite dimensional spaces can be substantially

different. The functional setting generalizes the vector-based analysis in the preceding subsection, but its infinite-dimensional nature allows for a richer specification of perturbation constraints that align with the geometry of functional data.

These findings offer several key insights into the behavior and potential of VAT when applied to functional data, such as time series. A critical implication of these results lies in the role of the space E (or equivalently, the norm $\|\cdot\|_E$), which defines the constraint of the perturbation r . It is known that the stronger norm of E (thus a “smaller” space) leads to a weaker norm of E^* (thus a “bigger” space), which can be seen from the duality between $\|\cdot\|_\Sigma$ and $\|\cdot\|_{\Sigma^{-1}}$. We will use the Sobolev spaces to show a concrete example in the next section. Consequently, the result highlights that, if we want to impose a stronger norm on the sensitivity of the method, we should, in contrast, relax the norm of the perturbation.

Moreover, this perspective also has practical implications for designing VAT-based training procedures. The choice of E and thus E^* determines the directions and structure of perturbations against which the model is made robust. For time series data, where the inputs are functions over a time interval T , this flexibility is particularly valuable. Time series often exhibit structural properties such as smoothness, periodicity, or specific temporal dependencies. By selecting a space that captures these characteristics—for instance, a Sobolev space that penalizes high-frequency oscillations through derivative-based norms—the VAT loss can be tailored to prioritize robustness under a particular sense. This could be especially beneficial in applications where high-frequency components are considered noise, which encourages the model to focus on the underlying smooth signal.

3.3 Structured Adversarial Perturbations via Sobolev Spaces

Sobolev spaces [1] are a powerful tool in functional analysis, particularly useful for analyzing the smoothness and regularity of functions. Using Sobolev spaces, we can define a structured norm for the f-VAT constraint that captures the smoothness properties of the perturbations. In this subsection, we will briefly introduce the Sobolev spaces, their duality, and practical ways to compute their norms with discrete points. We refer the readers to Appendix B in the appendix for more details.

For a positive integer s , the Sobolev space $H^s(T)$ comprises functions in $L^2(T)$ whose weak derivatives up to order s are also in $L^2(T)$. The weak derivative generalizes differentiation to functions lacking classical smoothness, enabling control over their regularity. A function with classical continuous derivatives up to order s is in $H^s(T)$, but the converse is not necessarily true. The norm in a Sobolev space is simply the sum of the $L^2(T)$ norms of all derivatives up to order s and the inner product can be similarly defined, which make $H^s(T)$ a Hilbert space.

More broadly, $H^\alpha(T)$ can be defined for any real $\alpha \in \mathbb{R}$. Non-integer α captures fractional smoothness, while negative α encompasses distributions rougher than $L^2(T)$ functions. Moreover, it is known that for $s > 0$, H^s is densely embedded into $L^2(T)$, while $L^2(T)$ is dense in $H^{-s}(T)$.

The duality of Sobolev spaces can be easily identified with the index s . The dual of $H^\alpha(T)$ is $H^{-\alpha}(T)$ under the $L^2(T)$ inner product. This duality enables us to frame the VAT loss using the dual norm $\|\cdot\|_{H^{-s}}$, assessing model sensitivity to perturbations constrained by $\|\cdot\|_{H^s}$ for $s > 0$.

Computing the Norm with Discrete Observations. In practical settings, time series are discrete vectors in $\mathbb{R}^N$. For simplicity, we assume that the time series is sampled at N equidistant points, but the results can be extended to non-equidistant points with minor adjustments. Then, we can approximate the Sobolev norms using finite difference operators [9]. Define the first-order difference matrix $D_1 \in \mathbb{R}^{N \times N}$ with $(D_1 r)_i = r_{i+1} - r_i$ and higher-order differences $D_k = D_1^k$. For integer k , the squared H^k norm of k with discrete observations $r_N \in \mathbb{R}^N$ can be approximated as:

$$\|r\|_{H^k}^2 \approx r_N^\top A_{k,N} r_N, \quad \text{where} \quad A_{k,N} = I_N + D_1^\top D_1 + D_2^\top D_2 + \cdots + D_k^\top D_k,$$

where I_N is the $N \times N$ identity matrix. This sums the ℓ_2 -norm of r and its discrete derivatives up to order k .

However, we actually will use the H^{-s} norm for the perturbation, so the above formulation is not directly applicable. For general $\alpha \in \mathbb{R}$, we will use the spectral method [9] instead. It is known that the Sobolev space $H^\alpha(T)$ can be represented via the fractional power of the Laplacian operator. We have $H^\alpha(\mathbb{R}^d) = \{f = (I - \Delta)^{-\alpha/2} u : u \in L^2(\mathbb{R}^d)\}$ and $\|f\|_{H^\alpha}^2 = \|(I - \Delta)^{\alpha/2} f\|_{L^2(\mathbb{R}^d)}^2$.

Therefore, let us take the discrete negative Laplacian matrix

$$L_N = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix}.$$

We can compute the H^α norm by

$$\|r\|_{H^\alpha}^2 \approx r_N^\top (I_N + L_N)^\alpha r_N, \quad \alpha \in \mathbb{R}, \quad (11)$$

where $(I_N + L_N)^\alpha$ is defined via the spectral theorem: if $I_N + L_N = Q\Lambda Q^\top$ with diagonal Λ , then $(I_N + L_N)^\alpha = Q\Lambda^\alpha Q^\top$, where Λ^α takes the exponent of each diagonal entry. This formulation is also valid for non-integer or negative s , where we note that $(I_N + L_N)$ is positive definite and thus the diagonal entries of Λ are positive. This gives us practical ways to compute the Sobolev norms with discrete observations.

3.4 The Algorithm

Let us now summarize the f-VAT procedure. Overall, we apply a mini-batch gradient descent to minimize the total loss. In practice, since the analytic expression of optimal perturbation r^* is intractable in Eq. (2), we approximate it using gradient ascent. We illustrate each step of updating θ in Algorithm 1, where we recall the labeled loss $\mathcal{L}_0(\mathcal{D}^l; f_\theta)$ and LDS in Eq. (1). According to our theory in Theorem 3.3 and the duality of Sobolev spaces, we use H^{-s} for the norm of adversarial perturbation r to impose a smoothness constraint on the functional model. In this paper, we mainly use $s = 2$ to penalize (weak) derivative up to the second order of the model, empirically verified by ablation studies in Fig. 1.

Algorithm 1 Functional Virtual Adversarial Training Step

- 1: **Input:** Data batch $\mathcal{D}, \mathcal{D}^l$, model f_θ , order of the Sobolev norm $s \geq 0$, radius ϵ , adversarial iterations L , learning rate η .
 - 2: **for** each sample $X_i \in \mathcal{D}$ **do** $\triangleright$ Approximate r_i^*
 - 3: Randomly initialize perturbation vector r_i over $\|r_i\|_{H^{-s}} \leq \epsilon$.
 - 4: **for** $\ell = 1 \rightarrow L$ **do**
 - 5: Gradient ascent $r_i \leftarrow r_i + \eta \nabla_{r_i} \text{LDS}(X_i, r_i; f_\theta)$
 - 6: Normalize $r_i \leftarrow \epsilon \frac{r_i}{\|r_i\|_{H^{-s}}}$.
 - 7: **end for**
 - 8: **end for**
 - 9: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$, where $\mathcal{L}(\theta) = \mathcal{L}_0(\mathcal{D}^l; f_\theta) + \frac{1}{|\mathcal{D}|} \sum_{X_i \in \mathcal{D}} \text{LDS}(X_i, r_i; f_\theta)$
-

4 Experiments

4.1 Experimental Settings

We use dozens of publicly available datasets from the UCR and UEA repositories [11], including the representative univariate dataset (i.e., CricketX, UWave, and InsectWing) and the multivariate dataset (i.e., SelfReg, NATOPS, and Heartbeat) in [19]. These representative datasets are from difficult to easy, and widely-used in semi-supervised time series classification [19, 21]. Additionally, we construct more empirical results on several large-scale China Securities Index (CSI) datasets (i.e., CSI 50 and 500 futures) spanning from 2020 to 2023 for predicting directions (upward or downward) of futures prices [44, 33]. The dataset collects records spanning from 2020 to 2022. Each time step contains bid/ask prices and corresponding volumes. Following [19], each dataset is split into train (60%), valid (20%), and test set (20%). We rescale each dataset into the range $[0, 1]$ for numerical stability. We refer readers to Appendix C for more details.

4.2 Performance Evaluation

Semi-Supervised Performance Following [19], we compare f-VAT with six deep learning-based methods, including recent SemiTime [19] and TapNet [43]. The baseline SupL represents the deep model trained only on the labeled dataset. We adopt the baseline results to Table 1 from the original work [19, 25, 43]. Following [40], we average the performance of each method over five runs with different random seeds and train/valid/test splits. Table 1 reports the mean and standard error from multiple runs.

Table 1: The semi-supervised classification accuracy (%) with standard deviation across six real-world datasets. Best performance in boldface.

Dataset	Ratio	SupL	PI	MTL	SemiTime	TapNet	VAT	f-VAT
CricketX	10%	44.88 $\pm$ 0.51	38.87 $\pm$ 2.26	40.94 $\pm$ 1.97	44.88 $\pm$ 3.13	39.42 $\pm$ 0.82	42.85 $\pm$ 3.97	49.18 $\pm$ 1.96
	20%	51.61 $\pm$ 0.45	44.44 $\pm$ 2.91	50.12 $\pm$ 1.22	51.61 $\pm$ 0.66	51.41 $\pm$ 0.31	49.14 $\pm$ 0.50	57.91 $\pm$ 3.58
	40%	58.71 $\pm$ 0.46	53.39 $\pm$ 2.18	55.10 $\pm$ 1.12	58.71 $\pm$ 2.78	58.97 $\pm$ 0.72	58.63 $\pm$ 0.50	68.39 $\pm$ 2.25
UWave	10%	81.46 $\pm$ 0.18	81.53 $\pm$ 0.54	76.35 $\pm$ 0.56	81.46 $\pm$ 0.60	82.34 $\pm$ 0.58	94.41 $\pm$ 0.09	94.82 $\pm$ 0.39
	20%	84.57 $\pm$ 0.87	81.66 $\pm$ 0.74	81.77 $\pm$ 0.94	84.57 $\pm$ 0.49	86.35 $\pm$ 0.43	95.53 $\pm$ 0.31	96.45 $\pm$ 0.27
	40%	86.91 $\pm$ 0.98	86.45 $\pm$ 1.20	86.91 $\pm$ 0.68	86.91 $\pm$ 0.47	89.24 $\pm$ 0.69	94.76 $\pm$ 0.54	97.23 $\pm$ 0.43
InsectWing	10%	54.96 $\pm$ 1.25	43.16 $\pm$ 3.20	50.45 $\pm$ 1.01	54.96 $\pm$ 1.61	55.53 $\pm$ 1.18	55.49 $\pm$ 1.28	58.01 $\pm$ 1.12
	20%	59.01 $\pm$ 1.13	48.35 $\pm$ 0.81	56.43 $\pm$ 0.88	59.01 $\pm$ 1.56	60.36 $\pm$ 0.38	61.27 $\pm$ 0.19	61.28 $\pm$ 1.86
	40%	62.38 $\pm$ 1.39	55.32 $\pm$ 2.04	60.90 $\pm$ 0.87	62.38 $\pm$ 0.76	63.87 $\pm$ 1.41	63.48 $\pm$ 0.30	64.81 $\pm$ 1.15
SelfReg	10%	46.49 $\pm$ 2.01	50.44 $\pm$ 0.76	50.88 $\pm$ 2.01	49.68 $\pm$ 2.83	50.87 $\pm$ 3.31	53.12 $\pm$ 4.51	59.31 $\pm$ 3.06
	20%	52.44 $\pm$ 3.15	53.94 $\pm$ 2.63	52.19 $\pm$ 2.01	52.63 $\pm$ 1.31	54.39 $\pm$ 2.74	55.76 $\pm$ 0.35	61.60 $\pm$ 1.13
	40%	51.31 $\pm$ 3.48	55.69 $\pm$ 2.74	56.14 $\pm$ 2.01	49.56 $\pm$ 1.72	54.38 $\pm$ 0.76	53.47 $\pm$ 1.04	64.44 $\pm$ 3.13
NATOPS	10%	68.98 $\pm$ 2.89	75.83 $\pm$ 4.39	73.91 $\pm$ 3.73	68.52 $\pm$ 0.81	70.37 $\pm$ 7.12	82.38 $\pm$ 0.96	86.04 $\pm$ 1.41
	20%	81.02 $\pm$ 1.60	82.51 $\pm$ 1.25	82.41 $\pm$ 2.89	80.09 $\pm$ 2.12	77.77 $\pm$ 1.39	82.81 $\pm$ 0.52	86.25 $\pm$ 1.38
	40%	88.89 $\pm$ 2.78	88.27 $\pm$ 1.19	90.27 $\pm$ 1.39	87.49 $\pm$ 2.41	82.87 $\pm$ 2.12	90.15 $\pm$ 1.60	93.13 $\pm$ 0.15
Heartbeat	10%	67.08 $\pm$ 3.57	72.13 $\pm$ 1.99	71.61 $\pm$ 2.47	71.61 $\pm$ 1.71	72.84 $\pm$ 1.23	73.86 $\pm$ 0.59	76.25 $\pm$ 1.22
	20%	73.25 $\pm$ 0.71	72.01 $\pm$ 0.78	73.66 $\pm$ 0.71	74.49 $\pm$ 1.43	73.24 $\pm$ 1.88	71.59 $\pm$ 0.13	76.46 $\pm$ 1.06
	40%	67.08 $\pm$ 1.89	73.28 $\pm$ 1.53	73.61 $\pm$ 3.07	72.43 $\pm$ 3.11	73.66 $\pm$ 0.71	75.00 $\pm$ 0.11	77.28 $\pm$ 0.40

Table 1 shows that f-VAT consistently outperforms other competitive baselines across all real-world datasets reported in various label ratios. For instance, f-VAT achieves up to 9.42% performance improvements on CricketX and 8.30% on SelfReg with label ratio $\alpha = 0.4$, while in the label-scarce scenario $\alpha = 0.1$, f-VAT still achieves up to 6.19% performance improvements on SelfReg.

To further verify the superiority of f-VAT, we randomly sample 30 datasets from UCR/UEA datasets [16]. Then, we add a classical baseline meanTeacher [38] and a recent state-of-the-art Class-Aware Temporal and Contextual Contrasting (CA-TCC) [18] for fair comparison. Table 2 shows that f-VAT still consistently outperforms other competitive baselines in all settings. Due to the page limit, the details of more empirical results can be found in Appendix E.

Table 2: The average accuracy (%) and average rank under different label ratios.

Method	10%		20%		40%	
	AvgAcc	AvgRank	AvgAcc	AvgRank	AvgAcc	AvgRank
SupL	35.31	6.67	36.92	7.00	37.15	7.33
PI	53.09	3.93	55.16	4.40	63.60	4.47
MTL	45.19	5.70	45.72	5.87	46.11	6.80
meanTeacher	42.89	5.93	50.94	4.87	63.85	4.13
SemiTime	56.53	3.57	58.93	3.77	69.02	3.13
TapNet	58.67	3.70	60.41	3.40	70.28	3.17
CA-TCC	58.07	3.37	59.84	3.67	63.27	4.27
f-VAT	65.85	1.50	68.87	1.50	76.24	1.53

Additionally, we construct more empirical results on several large-scale China Securities Index (CSI) datasets (i.e., CSI 50 and 500 futures). Table 3 presents the performance of various semi-supervised methods across different label ratios, and shows that f-VAT significantly outperforms other competitive baselines on futures datasets, especially on more volatile CSI 500 futures. This is because f-VAT’s adversarial perturbations incorporating key temporal structure facilitate deep models to effectively use unlabeled samples to yield smoother predictive distribution with better generalization.

Table 3: The performance comparison on domestic futures datasets like CSI 50 and 500 futures. Best performance in boldface.

Futures	Ratio	SupL	PI	MTL	SemiTime	TapNet	CA-TCC	f-VAT
50	10%	40.05 ± 1.38	42.87 ± 1.45	53.94 ± 0.13	55.19 ± 0.77	54.62 ± 0.54	55.72 ± 1.16	58.64 ± 0.53
	20%	45.08 ± 1.45	47.26 ± 0.73	54.97 ± 0.61	56.69 ± 0.50	56.93 ± 1.28	58.21 ± 0.44	62.09 ± 0.26
	40%	50.69 ± 3.53	52.42 ± 1.21	56.97 ± 0.57	57.34 ± 0.20	59.75 ± 1.09	59.80 ± 1.32	64.78 ± 0.45
500	10%	34.23 ± 0.24	44.00 ± 0.06	39.53 ± 1.35	38.86 ± 2.04	40.53 ± 0.83	39.79 ± 1.54	43.77 ± 0.73
	20%	35.38 ± 1.06	46.57 ± 0.48	45.26 ± 0.29	46.04 ± 0.12	44.58 ± 1.70	45.05 ± 1.74	52.14 ± 0.25
	40%	43.85 ± 1.49	49.25 ± 0.55	47.66 ± 1.50	50.65 ± 1.05	51.30 ± 1.36	54.29 ± 0.87	58.66 ± 0.46

Fully-Supervised Performance The proposed f-VAT can be easily extended to fully supervised settings. We compare f-VAT with several competitive supervised learning methods. ED [3] is the classical one-nearest-neighbor classifier based on Euclidean distance. TapNet [43] and ShapeNet [26] are deep learning-based methods that leverage manual shapelet-based features for better representations. Additionally, we include two non-neural methods called ROCKET [17] and HiveCOTE [4]. The empirical results in Table 4 are taken from the original work [3, 43, 26]. We average the performance of each method over five runs with different random seeds, and report mean values with standard deviations.

Table 4: The performance comparison between f-VAT and other baselines in fully-supervised settings. We report the mean and standard deviation over five runs. Best performance in boldface.

Dataset	Hive-COTE	ROCKET	ED	TapNet	ShapeNet	VAT	f-VAT
CricketX	74.10 ± 0.03	76.10 ± 0.01	62.90 ± 0.14	66.20 ± 0.25	68.30 ± 0.51	68.54 ± 1.40	77.25 ± 0.94
UWave	92.10 ± 0.02	93.70 ± 0.04	88.10 ± 0.12	89.40 ± 0.69	90.60 ± 0.13	92.43 ± 0.47	97.75 ± 0.13
InsectWing	62.20 ± 0.01	64.70 ± 0.01	60.20 ± 0.13	67.30 ± 0.11	66.30 ± 0.02	70.01 ± 1.16	71.70 ± 0.55
SelfReg	51.60 ± 0.67	51.40 ± 0.59	48.30 ± 0.12	55.10 ± 0.26	57.80 ± 0.03	58.75 ± 1.25	60.21 ± 0.68
NATOPS	82.80 ± 0.32	88.50 ± 0.44	85.10 ± 0.18	93.90 ± 0.01	88.30 ± 0.03	87.58 ± 1.89	97.50 ± 0.51
Heartbeat	72.20 ± 0.52	71.70 ± 0.02	61.90 ± 0.09	72.10 ± 1.43	75.60 ± 0.02	76.08 ± 0.82	78.75 ± 0.41

Table 4 shows that f-VAT consistently outperforms other baselines with an average improvement of 2.4% in all settings, especially for datasets containing long-term trend structures (i.e., up to 4.05% improvement in UWave). Also, f-VAT exhibits lower variance than VAT, as the functional adversarial perturbation imposes stronger regularization on the model’s gradient sensitivity and improves the model’s stability in fully-supervised settings.

Altering Order of Sobolev Norm We further explore how the order s of Sobolev norm influences the performance of f-VAT. Fig. 1 illustrates that the performance of f-VAT in NATOPS with label ratio $\alpha = 0.4$ under different s , where the error bar reports the standard deviation over 5 runs. We observe that the model’s accuracy peaks at $s = 2$. This is because under-regularization ($s \leq 1$) or over-regularization ($s \geq 3$) both would degenerate predictive performance. Based on this, we choose $s = 2$ in the rest of experiments. Additionally, we evaluate the performance on more datasets in Table 5, where $s = 0$ reduces to the original VAT. Table 5 shows that setting $s = 2$ achieves the best performance in almost all settings. For high volatile time series data, setting high order $s = 3$ generates adversarial perturbations that preserve low-frequency temporal information, while for relatively stable time series data, the small order $s = 1$ allows the perturbation to flexibly explore input space. In practice, it suffices to evaluate $s \in \{1, 2, 3, 4\}$ based on the validation set and report the average performance over several runs.

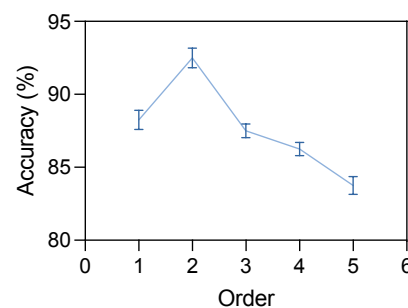


Figure 1: Various order of Sobolev norm.

Moreover, to verify the model-agnostic property of f-VAT, we compare the performance of different architectures trained by f-VAT in Appendix D. The empirical results suggest that Temporal Convolutional Network (TCN) [5] outperforms other architectures in most settings, so we adopt the eight-layer TCN as main architecture.

Table 5: The performance of f-VAT with different Sobolev norm order s .

s	CricketX	UWave	InsectWing	NATOPS	SelfReg
0	58.63 $\pm$ 0.50	94.76 $\pm$ 0.54	63.48 $\pm$ 0.30	90.15 $\pm$ 1.60	53.47 $\pm$ 1.04
1	59.91 $\pm$ 2.32	96.54 $\pm$ 0.67	66.70 $\pm$ 0.50	89.58 $\pm$ 0.12	56.16 $\pm$ 1.73
2	61.66 $\pm$ 2.33	97.16 $\pm$ 0.28	67.08 $\pm$ 0.86	93.13 $\pm$ 0.15	58.86 $\pm$ 0.35
3	60.44 $\pm$ 0.23	96.82 $\pm$ 0.16	64.10 $\pm$ 0.86	90.10 $\pm$ 0.65	51.39 $\pm$ 1.21
4	58.22 $\pm$ 3.39	96.71 $\pm$ 0.61	66.11 $\pm$ 0.82	87.51 $\pm$ 0.52	50.93 $\pm$ 0.12

4.3 Further Analysis

In this subsection, we conduct additional qualitative experiments to further analyze the behaviors of the deep model trained by f-VAT in the semi-supervised time series classification.

Perturbation Visualization Fig. 2 visualizes various virtual perturbations on NATOPS with label ratio $\alpha = 0.4$. The **color** line represents the original time series data, the **orange** line adds adversarial perturbation generated by the original VAT, and the **blue** line adds our proposed functional adversarial perturbation by f-VAT. As shown in Fig. 2, the functional adversarial perturbation closely aligns with the original sample across all key time steps and only adds subtle fluctuations to non-critical regions, while the original VAT’s perturbation introduces large spikes that disrupt underlying trend structure. These observations show that functional adversarial perturbations can preserve trend information without introducing anomalous patterns, facilitating deep models to achieve smoother predictive distributions and more stable convergence.

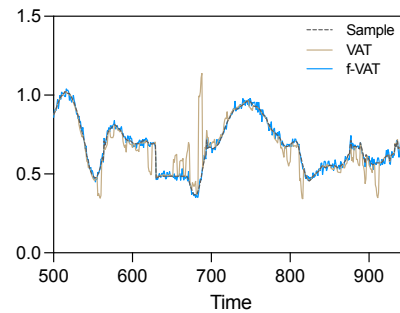


Figure 2: Visualization of perturbations.

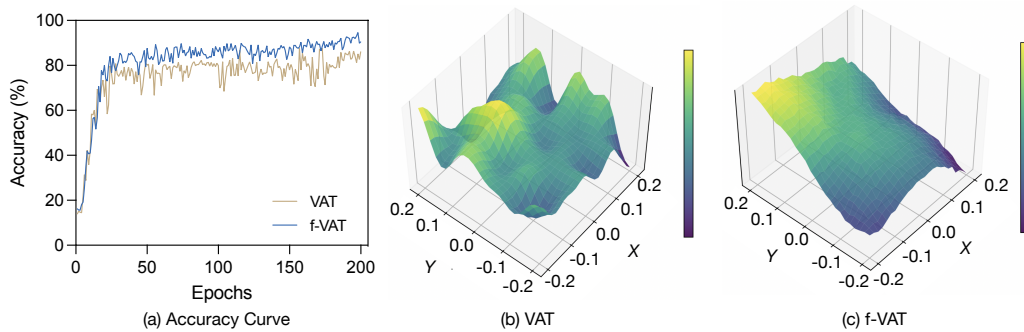


Figure 3: Visualization of performance curve (left) and the loss landscape of deep models trained by VAT (mid) and f-VAT (right) on the UWave dataset with label ratio $\alpha = 0.4$. The landscape of VAT contains many local minima, trapping the model in suboptimal solutions. The landscape of f-VAT is significantly smoother due to the adversarial perturbation in function space. Thus, f-VAT achieves faster and more stable convergence and better final results (left).

Loss Landscapes To validate the smoother predictive distribution offered by the functional adversarial perturbation in Sobolev space and demonstrate that f-VAT leads to easier optimization, we visualize the loss landscape of deep model trained by f-VAT and VAT using “filter normalization” [27] on NATOPS with label ratio $\alpha = 0.4$. From Fig. 3(mid), we observe that the loss landscape of VAT is highly chaotic, which causes the model to be easily trapped in suboptimal minima, leading to an unstable training curve. In contrast, deep models trained by f-VAT (as shown in Fig. 3 (right)) enjoy a significantly smoother loss landscape, where the functional adversarial perturbation in the Sobolev space preserves low-frequency trend information to provide more consistent gradient directions.

Consequently, f-VAT facilitates deep models for faster and more stable convergence with better final performance, as demonstrated in Fig. 3 (left).

Feature Importance Analysis To further analyze the behaviors of deep models trained by f-VAT in semi-supervised settings, Fig. 4 visualizes the gradient-based feature importance map [34] on NATOPS with different label ratios. The sample corresponds to “Fold wings” action, whose feature importance significantly changes within the time interval [25, 35]. These observations show that as supervision signals increase, deep model trained by f-VAT captures essential “shapelets” (i.e., “elbow’s movement to the right”) [23] and reduces its reliance on irrelevant features (i.e., feature 2 “shoulder rotation”), which aligns with existing knowledge that “Fold wings” action primarily relies on “elbow movement” with less reliance on “shoulder notation” [37]. This qualitative analysis validates the importance of capturing functional trend information within time series data and demonstrates that f-VAT is a highly effective and efficient method for this purpose.

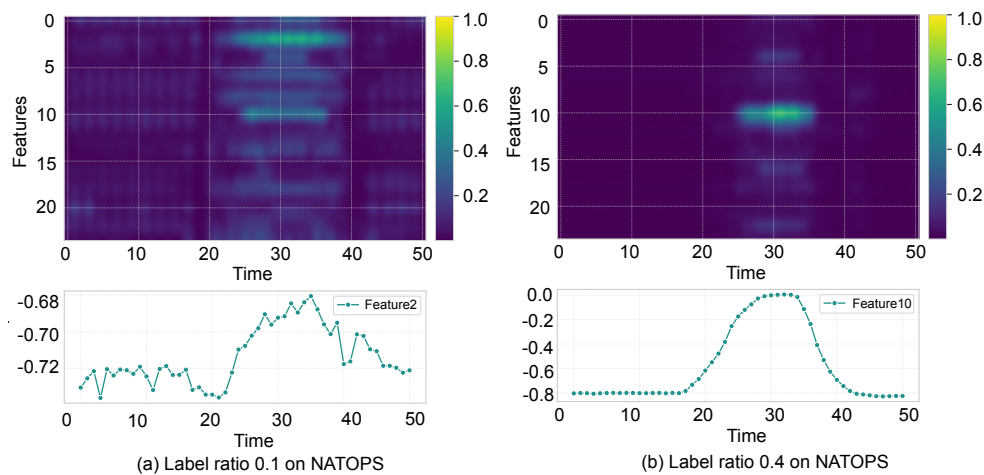


Figure 4: As the label ratio increases, the deep model trained by f-VAT ignores less important features (e.g., feature 2) and captures more critical regions (e.g., feature 10).

5 Limitations and Future Directions

In this paper, we propose functional Virtual Adversarial Training (f-VAT), which incorporates the functional structure of data into perturbations. Considering semi-supervised time series classification, we provide both theoretical insights and extensive empirical results showing the superiority of the proposed f-VAT method. We believe that our f-VAT method can serve as a general framework for semi-supervised learning methods with functional data.

Nevertheless, there are still some limitations in this work. We follow the implementations of representative semi-supervised time series classification settings, which primarily train backbone models from scratch, rather than fine-tuning them based on pretrained weights. We consider examining the effectiveness of pre-trained backbones as a potential future work.

Another future direction is to explore and design different adversarial perturbation norm for other functional models. For example, for data with known seasonality or periodic patterns, we can design a norm that additionally penalizes the non-periodicity of the perturbation to align with the periodic patterns, which could be done using the Fourier domain decomposition. We believe that investigating additional structure of the time series would further enhance the performance of the f-VAT method. Also, if the functional data lie on a manifold, one can introduce norm induced by the manifold structure to better align the perturbations. These further exploration would shed the light on the development of semi-supervised learning methods.

Acknowledgments and Disclosure of Funding

We would like to thank the anonymous reviewers and area chairs for their valuable comments and suggestions.

References

- [1] Robert A Adams and John JF Fournier. *Sobolev spaces*, volume 140. Elsevier, 2003.
- [2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.
- [3] Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.
- [4] Anthony Bagnall, Michael Flynn, James Large, Jason Lines, and Matthew Middlehurst. A tale of two toolkits, report the third: on the usage and performance of hive-cote v1. 0. *Advanced Analytics and Learning on Temporal Data*, 2020.
- [5] 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.
- [6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- [7] Niels Birbaumer, Nimr Ghanayim, Thilo Hinterberger, Iver Iversen, Boris Kotchoubey, Andrea Kübler, Juri Perlmouter, Edward Taub, and Herta Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.
- [8] George EP Box and Gwilym M Jenkins. Time series analysis, forecasting and control. *Holden-Day*, 1976.
- [9] Claudio Canuto, M. Youssuff Hussaini, Alfio Quarteroni, and Thomas A. Zang. *Spectral Methods: Fundamentals in Single Domains*. Scientific Computation. Springer Berlin Heidelberg, 2006.
- [10] Mingcai Chen, Hao Cheng, Yuntao Du, Ming Xu, Wenyu Jiang, and Chongjun Wang. Two wrongs don't make a right: Combating confirmation bias in learning with label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14765–14773, 2023.
- [11] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data/.
- [12] Yanping Chen, Adena Why, Gustavo Batista, Agenor Mafra-Neto, and Eamonn Keogh. Flying insect classification with inexpensive sensors. *Journal of Insect Behavior*, 27(5):657–677, 2014.
- [13] Constantin Christof. On the stability properties and the optimization landscape of training problems with squared loss for neural networks and general nonlinear conic approximation schemes. *Journal of Machine Learning Research*, 22(263):1–77, 2021.
- [14] Mooi Choo Chuah and Fen Fu. Ecg anomaly detection via time series analysis. In *Frontiers of High Performance Computing and Networking ISPA 2007 Workshops: ISPA 2007 International Workshops SSDSN, UPWN, WISH, SGC, ParDMCom, HiPCoMB, and IST-AWSN Niagara Falls, Canada, August 28-September 1, 2007 Proceedings* 5, pages 123–135. Springer, 2007.
- [15] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

- [16] 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.
- [17] Angus Dempster, François Petitjean, and Geoffrey I Webb. Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020.
- [18] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. Self-supervised contrastive representation learning for semi-supervised time-series classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15604–15618, 2023.
- [19] Haoyi Fan, Fengbin Zhang, Ruidong Wang, Xunhua Huang, and Zuoyong Li. Semi-supervised time series classification by temporal relation prediction. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 3545–3549. IEEE, 2021.
- [20] Nehla Ghouaiel, Pierre-François Marteau, and Marc Dupont. Continuous pattern detection and recognition in stream-a benchmark for online gesture recognition. *International Journal of Applied Pattern Recognition*, 4(2):146–160, 2017.
- [21] Jann Goschenhofer, Rasmus Hvingelby, David Rügamer, Janek Thomas, Moritz Wagner, and Bernd Bischl. Deep semi-supervised learning for time series classification. *arXiv preprint arXiv:2102.03622*, 2021.
- [22] James D Hamilton. *Time series analysis*. Princeton university press, 2020.
- [23] Jon Hills, Jason Lines, Edgaras Baranauskas, James Mapp, and Anthony Bagnall. Classification of time series by shapelet transformation. *Data mining and knowledge discovery*, 28:851–881, 2014.
- [24] Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 1 edition, 2015.
- [25] Shayan Jawed, Josif Grabocka, and Lars Schmidt-Thieme. Self-supervised learning for semi-supervised time series classification. *Advances in Knowledge Discovery and Data Mining*, 12084:499, 2020.
- [26] Guozhong Li, Byron Choi, Jianliang Xu, Sourav S Bhowmick, Kwok-Pan Chun, and Grace Lai-Hung Wong. Shapenet: A shapelet-neural network approach for multivariate time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8375–8383, 2021.
- [27] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.
- [28] Linyang Li and Xipeng Qiu. Token-aware virtual adversarial training in natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8410–8418, 2021.
- [29] Chengyu Liu, David Springer, Qiao Li, Benjamin Moody, Ricardo Abad Juan, Francisco J Chorro, Francisco Castells, José Millet Roig, Ikaro Silva, Alistair EW Johnson, et al. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12):2181, 2016.
- [30] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. uwave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5(6):657–675, 2009.
- [31] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

- [32] Abdullah Mueen, Eamonn Keogh, and Neal Young. Logical-shapelets: an expressive primitive for time series classification. In *International Conference on Knowledge Discovery and Data Mining*, pages 1154–1162, 2011.
- [33] Qingyi Pan, Suyu Sun, Pei Yang, and Jingyi Zhang. Futuresnet: Capturing patterns of price fluctuations in domestic futures trading. *Electronics*, 13(22):4482, 2024.
- [34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [35] Hussein Sharadga, Shima Hajimirza, and Robert S Balog. Time series forecasting of solar power generation for large-scale photovoltaic plants. *Renewable Energy*, 150:797–807, 2020.
- [36] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.
- [37] Yale Song, David Demirdjian, and Randall Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 500–506. IEEE, 2011.
- [38] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.
- [39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [40] Haishuai Wang, Qin Zhang, Jia Wu, Shirui Pan, and Yixin Chen. Time series feature learning with labeled and unlabeled data. *Pattern Recognition*, 89:55–66, 2019.
- [41] Li Wei and Eamonn Keogh. Semi-supervised time series classification. In *International Conference on Knowledge Discovery and Data Mining*, pages 748–753, 2006.
- [42] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*, 2020.
- [43] Xuchao Zhang, Yifeng Gao, Jessica Lin, and Chang-Tien Lu. Tapnet: Multivariate time series classification with attentional prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6845–6852, 2020.
- [44] Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, 67(11):3001–3012, 2019.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The main claims made in the abstract and introduction accurately reflect our paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Our paper discusses the limitations of the work in Section 5 Limitations and Futures Directions.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Our paper provides a full set of assumptions and a complete (and correct) proof in Appendix A and Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Our paper fully discloses the information needed to reproduce the main experimental results in Section 4.1 and Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include our codes and instructions to reproduce our main experimental results in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Our paper specifies the training and test details in Section 4.1 and Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the mean and standard error over five runs in Table 1, Table 4 and Fig. 1 within Section 4.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper provides sufficient information on the computer resources in Appendix C.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducts with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The paper is essentially fundamental research in machine learning, without negative social impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed. There is no societal impact of the work performed
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Duality Characterization of VAT Loss in General Spaces

Let us first introduce some preliminaries on the duality with respect to L^2 . Let T be an interval in $\mathbb{R}$ and we denote by $L^2(T)$ the space of square-integrable functions on T with the inner product

$$\langle u, v \rangle_{L^2(T)} = \int_T u(x)v(x)dx.$$

Let E be a Banach space with norm $\|\cdot\|_E$. Let E^* be the dual space of E , which consists of all continuous linear functionals on E . For $f \in E^*$ and $u \in E$, we denote by $\langle f, u \rangle_{E^*, E} = f(u)$ the duality pairing between f and u . Suppose that E is either densely embedded in $L^2(T)$ or $L^2(T)$ is densely embedded in E . We will discuss the relationship between the dual space E^* and the space $L^2(T)$. We denote by $\|\cdot\|$ the operator norm.

The case $E \hookrightarrow L^2$. Let $i_{E \rightarrow L^2} : E \rightarrow L^2$ be the embedding. Suppose that $\|u\|_{L^2} \leq K\|u\|_E$. For $v \in L^2$ and $u \in E$, we define the functional

$$\phi_v(u) = \langle i_{E \rightarrow L^2}(u), v \rangle_{L^2}.$$

Then, we find that

$$|\phi_v(u)| = |\langle i_{E \rightarrow L^2}(u), v \rangle_{L^2}| \leq \|i_{E \rightarrow L^2}(u)\|_{L^2} \|v\|_{L^2} \leq \|i_{E \rightarrow L^2}\| \|u\|_E \|v\|_{L^2},$$

so

$$\|\phi_v\|_{E^*} = \sup_{\|u\|_E \leq 1} |\phi_v(u)| \leq \|i_{E \rightarrow L^2}\| \|v\|_{L^2}.$$

This shows that $\phi_v \in E^*$. Moreover, suppose $\phi_v \in E^*$ is identically zero. Then, for all $u \in E$, we have $\phi_v(u) = \langle i_{E \rightarrow L^2}(u), v \rangle_{L^2} = 0$. But since the range of $i_{E \rightarrow L^2}$ is dense in L^2 , we must have $v = 0$. Therefore, $v \mapsto \phi_v$ defines an embedding $L^2 \hookrightarrow E^*$, which we denote by $i_{L^2 \rightarrow E^*}$. In addition, we also have

$$\langle v, i_{E \rightarrow L^2}(u) \rangle_{L^2} = \langle i_{L^2 \rightarrow E^*}(v), u \rangle_{E^*, E}, \quad \text{namely, } i_{E \rightarrow L^2}^* = i_{L^2 \rightarrow E^*}.$$

If we view the embeddings as inclusions, we have $E \subset L^2 \subset E^*$.

The case $L^2 \hookrightarrow E$. Let $i_{L^2 \rightarrow E} : L^2 \rightarrow E$ be the embedding. For an element $\phi \in E^*$, $\phi \circ i_{L^2 \rightarrow E}$ also defines a continuous functional on L^2 . Consequently, from the Riesz representation theorem, there exists a unique element $v_\phi \in L^2$ such that $(\phi \circ i_{L^2 \rightarrow E})(u) = \langle u, v_\phi \rangle_{L^2}$. Moreover,

$$\|v_\phi\|_{L^2} = \sup_{\|u\|_{L^2} \leq 1} |(\phi \circ i_{L^2 \rightarrow E})(u)| = \sup_{\|u\|_{L^2} \leq 1} |\phi(i_{L^2 \rightarrow E}(u))| \leq \|\phi\|_{E^*} \|i_{L^2 \rightarrow E}\| \|u\|_{L^2}.$$

Therefore, the mapping $\phi \mapsto v_\phi$ is a bounded linear operator from E^* to L^2 , which we denote by $i_{E^* \rightarrow L^2}$. Furthermore, if v_ϕ is identically zero, then ϕ must also be zero by the density of the range of $i_{L^2 \rightarrow E}$. Consequently, $i_{E^* \rightarrow L^2}$ is an embedding. Similarly, we have

$$\langle \phi, i_{L^2 \rightarrow E}(u) \rangle_{E^*, E} = \langle i_{E^* \rightarrow L^2}(\phi), u \rangle_{L^2}, \quad \text{namely, } i_{L^2 \rightarrow E}^* = i_{E^* \rightarrow L^2}.$$

Viewing the embeddings as inclusions, we have $E^* \subset L^2 \subset E$.

A.1 The Duality Characterization of VAT Loss

With the above notations, let us give a more rigorous statement of the theorem. When $E \hookrightarrow L^2$, we define the VAT loss as

$$\mathcal{L}^{\text{VA}}(x_0, \epsilon; f) = \sup_{r = i_{E \rightarrow L^2}(r_0) \in L^2 : \|r_0\|_E \leq \epsilon} |f(x_0) - f(x_0 + r)|^2, \quad (12)$$

while when $L^2 \hookrightarrow E$, we define the VAT loss as

$$\mathcal{L}^{\text{VA}}(x_0, \epsilon; f) = \sup_{r \in L^2 : \|i_{L^2 \rightarrow E}(r)\|_E \leq \epsilon} |f(x_0) - f(x_0 + r)|^2, \quad (13)$$

Let us give a more rigorous statement of the theorem.

Theorem A.1. Let $f(x) = \langle \beta, x \rangle_{L^2(T)}$ be a functional linear model with $\beta \in L^2$.

- For the case $E \hookrightarrow L^2$, the loss Eq. (13) is equivalent to

$$\mathcal{L}^{\text{VA}}(x_0, \epsilon; f) = \epsilon^2 \|i_{L^2 \rightarrow E^*}(\beta)\|_{E^*}^2. \quad (14)$$

- For the case $L^2 \hookrightarrow E$, the loss Eq. (12) is equivalent to

$$\mathcal{L}^{\text{VA}}(x_0, \epsilon; f) = \begin{cases} \epsilon^2 \|\beta_0\|_{E^*}^2, & \text{if } \beta = i_{E^* \rightarrow L^2}(\beta_0) \text{ for some } \beta_0 \in E^*, \\ \infty, & \text{otherwise.} \end{cases} \quad (15)$$

If we view the embeddings as inclusions, we unify the two cases as

$$\mathcal{L}^{\text{VA}}(x_0, \epsilon; f) = \epsilon^2 \|\beta\|_{E^*}^2. \quad (16)$$

Proof. Let E_0 be the set over which the supremum is taken in Eq. (13) and Eq. (12). Then, the VAT loss writes

$$\mathcal{L}^{\text{VA}}(x_0, \epsilon; f) = \sup_{r \in E_0} |\langle \beta, x_0 \rangle - \langle \beta, x_0 + r \rangle|^2 = \sup_{r \in E_0} \langle \beta, r \rangle_{L^2(T)}^2$$

Let us consider the two cases of separately.

The case $E \hookrightarrow L^2$. Using the relation between the dual spaces, we have

$$\begin{aligned} \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) &= \sup_{r = i_{E \rightarrow L^2}(r_0) \in L^2: \|r_0\|_E \leq \epsilon} \langle \beta, r \rangle_{L^2(T)}^2 \\ &= \sup_{r_0 \in E: \|r_0\|_E \leq \epsilon} \langle \beta, i_{E \rightarrow L^2}(r_0) \rangle_{L^2(T)}^2 \\ &= \sup_{r_0 \in E: \|r_0\|_E \leq \epsilon} \langle i_{L^2 \rightarrow E^*}(\beta), r_0 \rangle_{E^*, E}^2 \\ &= \epsilon^2 \|i_{L^2 \rightarrow E^*}(\beta)\|_{E^*}^2. \end{aligned}$$

The case $L^2 \hookrightarrow E$. In this case, we have

$$\mathcal{L}^{\text{VA}}(x_0, \epsilon; f) = \sup_{r \in L^2: \|i_{L^2 \rightarrow E}(r)\|_E \leq \epsilon} \langle \beta, r \rangle_{L^2(T)}^2.$$

If $\beta = i_{E^* \rightarrow L^2}(\beta_0)$ for some $\beta_0 \in E^*$, we have

$$\begin{aligned} \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) &= \sup_{r \in L^2: \|i_{L^2 \rightarrow E}(r)\|_E \leq \epsilon} \langle i_{E^* \rightarrow L^2}(\beta_0), r \rangle_{L^2(T)}^2 \\ &= \sup_{r \in L^2: \|i_{L^2 \rightarrow E}(r)\|_E \leq \epsilon} \langle \beta_0, i_{L^2 \rightarrow E}(r) \rangle_{E^*, E}^2 \\ &= \sup_{r_1 \in E: \|r_1\|_E \leq \epsilon} \langle \beta_0, r_1 \rangle_{E^*, E}^2 \\ &= \epsilon^2 \|\beta_0\|_{E^*}^2. \end{aligned}$$

On the other hand, if $\mathcal{L}^{\text{VA}}(x_0, \epsilon; f)$ is finite, then we can define a function on $\text{Ran } i_{L^2 \rightarrow E}$ that

$$\phi(r_1) = \langle \beta, r \rangle_{L^2(T)}, \quad r_1 = i_{L^2 \rightarrow E}(r),$$

which is well-defined since $i_{L^2 \rightarrow E}$ is injective. It is easy to verify that ϕ is linear. Moreover,

$$\epsilon^2 \|\phi\|_{E^*}^2 = \sup_{r_1 \in E: \|r_1\|_E \leq \epsilon} |\phi(r_1)|^2 = \sup_{r \in L^2: \|i_{L^2 \rightarrow E}(r)\|_E \leq \epsilon} \langle \beta, r \rangle_{L^2(T)}^2 = \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) < \infty.$$

Hence, we have $\phi \in E^*$. Moreover, the definition of ϕ and the duality relation gives

$$\langle \beta, r \rangle_{L^2(T)} = \phi(r_1) = \langle \phi, r_1 \rangle_{E^*, E} = \langle \phi, i_{L^2 \rightarrow E}(r) \rangle_{E^*, E} = \langle i_{E^* \rightarrow L^2}(\phi), r \rangle_{L^2(T)},$$

which implies that $\beta = i_{E^* \rightarrow L^2}(\phi)$. Consequently, $\mathcal{L}^{\text{VA}}(x_0, \epsilon; f) < \infty$ if and only if $\beta \in \text{Ran } i_{E^* \rightarrow L^2}$ and we conclude the theorem. □

A.2 Proof of Theorem 3.4

Let us first recall that Fréchet differentiability of f at x_0 means that there exists a continuous linear functional $\nabla f(x_0) \in L^2(T)$ such that

$$f(x_0 + r) = f(x_0) + \langle \nabla f(x_0), r \rangle_{L^2(T)} + o(\|r\|_{L^2(T)}) \quad \text{as } \|r\|_{L^2(T)} \rightarrow 0.$$

Moreover, if $L^2 \rightarrow E$, we say that the Fréchet differentiability holds under the E -norm if $\nabla f(x_0) = i_{E^* \rightarrow L^2}(\beta_0)$ for some $\beta_0 \in E^*$ and for any $r \in L^2$, we have

$$f(x_0 + r) = f(x_0) + \langle \nabla f(x_0), r \rangle_{L^2(T)} + o(\|i_{L^2 \rightarrow E}(r)\|_E) \quad \text{as } \|r\|_E \rightarrow 0.$$

First, we give a rigorous version of Theorem 3.4 in the following.

Theorem A.2 (Theorem 3.4 restated). *Let $f : L^2(T) \rightarrow \mathbb{R}$ be Fréchet differentiable at $x_0 \in L^2(T)$ with gradient $\nabla f(x_0) \in L^2(T)$.*

- If $E \hookrightarrow L^2(T)$, then

$$\lim_{\epsilon \rightarrow 0^+} \epsilon^{-2} \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) = \|i_{L^2 \rightarrow E^*}(\nabla f(x_0))\|_{E^*}^2.$$

- If $L^2(T) \hookrightarrow E$, assume further that $\nabla f(x_0) = i_{E^* \rightarrow L^2}(\beta_0)$ for some $\beta_0 \in E^*$ and the Fréchet differentiability holds under E -norm, we have

$$\lim_{\epsilon \rightarrow 0^+} \epsilon^{-2} \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) = \|\beta_0\|_{E^*}^2.$$

Proof. We consider the two cases separately.

The case $E \hookrightarrow L^2(T)$ In this setting, the dual embedding $i_{L^2 \rightarrow E^*} : L^2(T) \rightarrow E^*$ satisfies

$$\langle i_{L^2 \rightarrow E^*}(v), u \rangle_{E^*, E} = \langle v, u \rangle_{L^2(T)} \quad \text{for all } v \in L^2(T), u \in E.$$

Combining it with the Fréchet differentiability of f at x_0 , If $r = i_{E \rightarrow L^2}(r_0)$, we have

$$\begin{aligned} f(x_0 + r) &= f(x_0) + \langle \nabla f(x_0), r \rangle_{L^2(T)} + o(\|r\|_{L^2(T)}) \\ &= f(x_0) + \langle i_{L^2 \rightarrow E^*}(\nabla f(x_0)), r_0 \rangle_{E^*, E} + o(\|i_{E \rightarrow L^2}(r_0)\|_{L^2(T)}) \\ &= f(x_0) + \langle i_{L^2 \rightarrow E^*}(\nabla f(x_0)), r_0 \rangle_{E^*, E} + o(\|r_0\|_E), \end{aligned}$$

Therefore, recalling that the VAT loss is given by Eq. (12), we have

$$\begin{aligned} \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) &= \sup_{r_0 \in E: \|r_0\|_E \leq \epsilon} \left(\langle i_{L^2 \rightarrow E^*}(\nabla f(x_0)), r_0 \rangle_{E^*, E} + o(\|r_0\|_E) \right)^2 \\ &= \sup_{r_0 \in E: \|r_0\|_E \leq \epsilon} \left(\langle i_{L^2 \rightarrow E^*}(\nabla f(x_0)), r_0 \rangle_{E^*, E} + o(\epsilon) \right)^2 \end{aligned}$$

Consequently,

$$\begin{aligned} \epsilon^{-2} \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) &= \sup_{r_0 \in E: \|r_0\|_E \leq \epsilon} \left(\epsilon^{-1} \langle i_{L^2 \rightarrow E^*}(\nabla f(x_0)), r_0 \rangle_{E^*, E} + o(1) \right)^2 \\ &= \sup_{s_0 \in E: \|s_0\|_E \leq 1} \left(\langle i_{L^2 \rightarrow E^*}(\nabla f(x_0)), s_0 \rangle_{E^*, E} + o(1) \right)^2 \\ &= \sup_{s_0 \in E: \|s_0\|_E \leq 1} \left(\langle i_{L^2 \rightarrow E^*}(\nabla f(x_0)), s_0 \rangle_{E^*, E} \right)^2 + o(1) \\ &= \|i_{L^2 \rightarrow E^*}(\nabla f(x_0))\|_{E^*}^2 + o(1). \end{aligned}$$

The case $L^2(T) \hookrightarrow E$ In this case, let us recall that the dual embedding $i_{E^* \rightarrow L^2} : E^* \rightarrow L^2(T)$ satisfies

$$\langle i_{E^* \rightarrow L^2}(\phi), u \rangle_{L^2(T)} = \langle \phi, u \rangle_{E^*, E} \quad \text{for all } \phi \in E^*, u \in L^2(T).$$

Thus, since $\nabla f(x_0) = i_{E^* \rightarrow L^2}(\beta_0)$ for some $\beta_0 \in E^*$, using the Fréchet differentiability of f at x_0 under the E -norm, we have

$$\begin{aligned} f(x_0 + r) &= f(x_0) + \langle \nabla f(x_0), r \rangle_{L^2(T)} + o(\|i_{L^2 \rightarrow E}(r)\|_E) \\ &= f(x_0) + \langle i_{E^* \rightarrow L^2}(\beta_0), r \rangle_{L^2(T)} + o(\|i_{L^2 \rightarrow E}(r)\|_E) \\ &= f(x_0) + \langle \beta_0, i_{L^2 \rightarrow E}(r) \rangle_{E^*, E} + o(\|i_{L^2 \rightarrow E}(r)\|_E). \end{aligned}$$

Recall the VAT loss is given by Eq. (13), we have

$$\begin{aligned} \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) &= \sup_{r \in L^2: \|i_{L^2 \rightarrow E}(r)\|_E \leq \epsilon} \left| \langle \beta_0, i_{L^2 \rightarrow E}(r) \rangle_{E^*, E} + o(\|i_{L^2 \rightarrow E}(r)\|_E) \right|^2 \\ &= \sup_{r_1 \in E, \|r_1\|_E \leq \epsilon} \left| \langle \beta_0, r_1 \rangle_{E^*, E} + o(\|r_1\|_E) \right|^2, \end{aligned}$$

where we use the density of $i_{L^2 \rightarrow E}$. Therefore,

$$\begin{aligned} \epsilon^{-2} \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) &= \sup_{r_1 \in E: \|r_1\|_E \leq \epsilon} \left(\epsilon^{-1} \langle \beta_0, r_1 \rangle_{E^*, E} + o(1) \right)^2 \\ &= \sup_{s_1 \in E: \|s_1\|_E \leq 1} \left(\langle \beta_0, s_1 \rangle_{E^*, E} + o(1) \right)^2 \\ &= \|\beta_0\|_{E^*}^2 + o(1). \end{aligned}$$

□

A.3 Proof of Proposition 3.2

Since $f(x)$ is continuously differentiable, we have

$$f(x_0) - f(x_0 + r) = \langle \nabla f(x_0), r \rangle + o(\|r\|_2).$$

Noticing that $\|r\|_{\Sigma} \leq \epsilon$ implies $\|r\|_2 \leq \epsilon \left\| \Sigma^{-\frac{1}{2}} \right\|_{\infty}$, we have

$$\begin{aligned} \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) &= \sup_{r: \|r\|_{\Sigma} \leq \epsilon} |\langle \beta, x_0 \rangle - \langle \beta, x_0 + r \rangle|^2 \\ &= \sup_{r: \|r\|_{\Sigma} \leq \epsilon} (\langle \nabla f(x_0), r \rangle + o(\|r\|_2))^2 \\ &= \sup_{r: \|r\|_{\Sigma} \leq \epsilon} (\langle \nabla f(x_0), r \rangle + o(\epsilon))^2. \end{aligned}$$

Therefore, when $\epsilon \rightarrow 0^+$, we obtain that

$$\begin{aligned} \epsilon^{-2} \mathcal{L}^{\text{VA}}(x_0, \epsilon; f) &= \sup_{r: \|r\|_{\Sigma} \leq \epsilon} \left(\epsilon^{-2} \langle \nabla f(x_0), r \rangle^2 + o(1) \right) \\ &\rightarrow \sup_{r: \|r\|_{\Sigma} \leq \epsilon} \epsilon^{-2} \langle \nabla f(x_0), r \rangle^2 \\ &= \|\nabla f(x_0)\|_{\Sigma^{-1}}^2. \end{aligned}$$

B Sobolev Spaces

Sobolev spaces are a cornerstone of functional analysis, providing a framework to quantify the smoothness of functions and distributions. This section outlines their construction, duality properties, and practical computation in discrete settings, which are particularly relevant for introducing dual loss in the VAT method. We refer to books [1, 9] for details.

B.1 Construction of Sobolev Spaces via Fourier Transforms

Sobolev spaces $H^s(\mathbb{R}^d)$ for $s \in \mathbb{R}$ are defined using Fourier transforms, offering a unified approach across all real smoothness indices. Let $\mathcal{S}(\mathbb{R}^d)$ denote the Schwartz space of smooth, rapidly decaying functions on $\mathbb{R}^d$, and $\mathcal{S}'(\mathbb{R}^d)$ its dual, the space of tempered distributions. Moreover, we can define the Fourier transformation $\hat{u}(\xi) = \mathcal{F}u(\xi) = \int_{\mathbb{R}^d} u(x)e^{-i\xi \cdot x} dx$ of $u \in \mathcal{S}(\mathbb{R}^d)$ and extend the Fourier transformation to $\mathcal{S}'(\mathbb{R}^d)$.

For $s \in \mathbb{R}$, the Sobolev space $H^s(\mathbb{R}^d)$ is defined as the space of tempered distributions u such that

$$(1 + |\xi|^2)^{s/2} |\hat{u}(\xi)| \in L^2(\mathbb{R}^d) \quad (17)$$

with the norm and inner product

$$\|u\|_{H^s} = \left(\int_{\mathbb{R}^d} (1 + |\xi|^2)^s |\hat{u}(\xi)|^2 d\xi \right)^{1/2}, \quad \langle u, v \rangle_{H^s} = \int_{\mathbb{R}^d} (1 + |\xi|^2)^s \hat{u}(\xi) \overline{\hat{v}(\xi)} d\xi. \quad (18)$$

Moreover, $H^s(\mathbb{R}^d)$ is also equivalent to the completion of $\mathcal{S}(\mathbb{R}^d)$ with respect to the H^s norm. The weight $(1 + |\xi|^2)^s$ modulates the contribution of different frequencies: for $s > 0$, it penalizes high-frequency components, enforcing smoothness; for $s < 0$, it emphasizes them, allowing rougher distributions; and for $s = 0$, it reduces to the $L^2(\mathbb{R}^d)$ norm, since $H^0(\mathbb{R}^d) = L^2(\mathbb{R}^d)$. It is well known that the Sobolev space $H^s(\mathbb{R}^d)$ is a Hilbert space under the inner-product $\langle \cdot, \cdot \rangle_{H^s}$.

For positive integer s , $H^s(\mathbb{R}^d)$ coincides with the space of L^2 functions whose weak derivatives up to order s are also in $L^2(\mathbb{R}^d)$. For negative s , $H^s(\mathbb{R}^d)$ includes distributions that lack the integrability of L^2 functions but are constrained by the decay of their Fourier coefficients.

In applications like time series, we often consider functions on a bounded domain, such as $T = [0, 1]$. For domains $\Omega \subseteq \mathbb{R}^d$, $H^s(\Omega)$ can be defined as the restriction of functions from $H^s(\mathbb{R}^d)$ to Ω , with the norm being the infimum of the $H^s(\mathbb{R}^d)$ norm over all extensions, or as the completion of $C^\infty(\Omega)$ with respect to the appropriate norm.

B.2 Duality and Embeddings of Sobolev Spaces

Sobolev spaces exhibit a rich duality structure, which is essential for understanding their properties and applications. For $s \in \mathbb{R}$, the dual space of $H^s(\mathbb{R}^d)$ with respect to the $L^2(\mathbb{R}^d)$ inner product is $H^{-s}(\mathbb{R}^d)$. That is, any continuous linear functional $\ell \in (H^s(\mathbb{R}^d))^*$ can be represented as:

$$\ell(u) = \langle v, u \rangle_{L^2} = \int_{\mathbb{R}^d} v(x) \overline{u(x)} dx, \quad (19)$$

for some $v \in H^{-s}(\mathbb{R}^d)$, with the norm equivalence:

$$\|\ell\|_{(H^s)^*} = \|v\|_{H^{-s}}. \quad (20)$$

This duality arises because the pairing $\langle v, u \rangle_{L^2}$ is well-defined when $u \in H^s$ and $v \in H^{-s}$, as the Fourier transform ensures that $(1 + |\xi|^2)^{-s/2} \hat{v}(\xi)$ and $(1 + |\xi|^2)^{s/2} \hat{u}(\xi)$ yield a product in $L^2(\mathbb{R}^d)$.

Sobolev spaces also satisfy embedding theorems, which relate them to other function spaces based on smoothness. A key result is the Sobolev embedding theorem: for $s > d/2$, $H^s(\mathbb{R}^d)$ embeds continuously into $C_b(\mathbb{R}^d)$, the space of bounded continuous functions, with:

$$\|u\|_{C_b} \leq C \|u\|_{H^s}, \quad (21)$$

for some constant C . This embedding implies that functions in H^s with sufficiently large s are not only continuous but also bounded, a property useful for ensuring regularity in optimization problems.

More generally, for $s_1 > s_2$, $H^{s_1}(\mathbb{R}^d) \hookrightarrow H^{s_2}(\mathbb{R}^d)$, reflecting that higher s corresponds to greater smoothness.

These properties—duality and embeddings—are critical in applications like Virtual Adversarial Training (VAT), where the dual norm $\|\nabla f(x_0)\|_{H^{-s}}$ quantifies model sensitivity to perturbations r constrained by $\|r\|_{H^s} \leq \epsilon$, and embeddings ensure perturbations maintain desirable regularity.

B.3 Fractional Powers of the Laplacian and Sobolev Norms

The fractional power of the Laplacian provides an operator-theoretic perspective on Sobolev spaces, unifying the definition of $H^s(\mathbb{R}^d)$ across all $s \in \mathbb{R}$. The Laplacian $-\Delta$ is a positive, self-adjoint operator on $L^2(\mathbb{R}^d)$, and its fractional power $(-\Delta)^\alpha$ for $\alpha \in \mathbb{R}$ is defined via the Fourier transform:

$$\mathcal{F}[(-\Delta)^\alpha u](\xi) = |\xi|^{2\alpha} \hat{u}(\xi),$$

where $\hat{u}(\xi) = \mathcal{F}u(\xi)$ is the Fourier transform of u . However, the Sobolev norm incorporates a shifted operator, $I - \Delta$, to ensure positivity and handle low frequencies effectively. Specifically, the fractional power $(I - \Delta)^{s/2}$ satisfies:

$$\mathcal{F}[(I - \Delta)^{s/2} u](\xi) = (1 + |\xi|^2)^{s/2} \hat{u}(\xi),$$

so that:

$$\|u\|_{H^s} = \|(I - \Delta)^{s/2} u\|_{L^2}.$$

This equivalence follows from:

$$\|(I - \Delta)^{s/2} u\|_{L^2}^2 = \int_{\mathbb{R}^d} |(1 + |\xi|^2)^{s/2} \hat{u}(\xi)|^2 d\xi = \int_{\mathbb{R}^d} (1 + |\xi|^2)^s |\hat{u}(\xi)|^2 d\xi = \|u\|_{H^s}^2.$$

For $s > 0$, $(I - \Delta)^{s/2}$ acts as a differential operator of order s , penalizing high-frequency oscillations and enforcing smoothness. For example, when $s = 2$, $(I - \Delta)u = u - \Delta u$, and the norm $\|u - \Delta u\|_{L^2}$ measures both the function and its second derivatives. For $s < 0$, $(I - \Delta)^{s/2}$ is a smoothing operator, and H^s includes distributions whose images under $(I - \Delta)^{-s/2}$ are in L^2 .

The dual norm in $H^{-s}(\mathbb{R}^d)$ relates to the inverse fractional power. For $v \in H^{-s}$, we have:

$$\|v\|_{H^{-s}} = \|(I - \Delta)^{-s/2} v\|_{L^2},$$

since:

$$\|v\|_{H^{-s}} = \sup_{\|u\|_{H^s} \leq 1} |\langle v, u \rangle_{L^2}| = \sup_{\|(I - \Delta)^{s/2} w\|_{L^2} \leq 1} |\langle v, (I - \Delta)^{s/2} w \rangle_{L^2}|,$$

and setting $w = (I - \Delta)^{-s/2} u$ yields the result via the self-adjointness of $I - \Delta$. This operator formulation is particularly useful in discrete settings, as it translates directly to matrix powers, as discussed previously.

The fractional Laplacian $(-\Delta)^s$ itself (without the identity shift) is also of interest, with norm:

$$\|(-\Delta)^{s/2} u\|_{L^2}^2 = \int_{\mathbb{R}^d} |\xi|^{2s} |\hat{u}(\xi)|^2 d\xi.$$

While this norm emphasizes derivative behavior alone, $I - \Delta$ ensures a baseline L^2 contribution, making H^s norms more robust for small s or low frequencies. Choosing between $(I - \Delta)^{s/2}$, $(-\Delta)^{s/2}$ or $(\alpha I - \Delta)^{s/2}$ for perturbation constraints can tailor the robustness profile: the former balances function magnitude and smoothness, while the latter focuses purely on derivative control.

B.4 Discrete Computation via Spectral Methods

In practical applications such as time series analysis, data are represented as discrete vectors in $\mathbb{R}^N$. To compute Sobolev norms in this discrete setting, we employ spectral methods that approximate the continuous operators while preserving their spectral properties, ensuring consistency with the operator perspective introduced in the discussion on fractional powers and Sobolev norms.

We recall that in the continuous case, the Sobolev norm of a function $u \in H^s(\mathbb{R}^d)$ is defined using the fractional power of the operator $I - \Delta$, where Δ is the Laplacian:

$$\|u\|_{H^s} = \|(I - \Delta)^{s/2} u\|_{L^2}.$$

This can be expressed in the Fourier domain as:

$$\|u\|_{H^s}^2 = \int_{\mathbb{R}^d} (1 + |\xi|^2)^s |\hat{u}(\xi)|^2 d\xi,$$

where $\hat{u}$ is the Fourier transform of u , and $1 + |\xi|^2$ arises from the eigenvalues of $I - \Delta$. In the discrete setting, we aim to approximate this norm for a vector $r \in \mathbb{R}^N$, representing a time series sampled at N equally spaced points. For simplicity, we assume periodic boundary conditions, which are common in time series analysis and allow for efficient computation via the Fast Fourier Transform (FFT).

B.4.1 Discrete Laplacian and Sobolev Norm

The discrete Laplacian $L_N \in \mathbb{R}^{n \times n}$ is constructed using second-order finite differences. For a uniform grid with n points, L_N is a tridiagonal matrix defined as:

$$L_N = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix}.$$

This matrix L_N approximates the negative Laplacian $-\Delta$ via twice difference, which is a positive semi-definite matrix. Thus, $I_N + L_N$, where I_N is the $N \times N$ identity matrix, approximates the operator $I - \Delta$. Consequently, the squared H^s norm of the discrete vector $r_N \in \mathbb{R}^n$ is then approximated as:

$$\|r\|_{H^s}^2 \approx r_N^\top (I_N + L_N)^s r_N.$$

B.4.2 Speeding up the Computation under Periodic Boundary Conditions

For many applications of time series, we assume periodic boundary conditions, which allow us to use the discrete Fourier transform (DFT) to compute the Sobolev norm efficiently. Under periodic boundary conditions, we take the discrete Laplacian as

$$\bar{L}_N = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & -1 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & -1 & 2 & -1 \\ -1 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix},$$

where we note the off-diagonal entries -1 at the corners (e.g., $L_{1,N} = -1$, $L_{N,1} = -1$) enforce periodicity. Then, this matrix $\bar{L}_N$ is a circulant matrix, which can be diagonalized using the discrete Fourier transform (DFT). The matrix can be diagonalized by the discrete Fourier transform matrix F , defined by:

$$F = (F_{jk})_{j,k=0}^{N-1}, \quad F_{jk} = \frac{1}{\sqrt{N}} e^{-i2\pi jk/N}, \quad j, k = 0, 1, \dots, N-1,$$

which is unitary (i.e., $F^* F = I_N$, with F^* being the conjugate transpose). Then, the spectral decomposition of $\bar{L}_N$ is given by:

$$\bar{L}_N = F \Lambda F^*,$$

where $\Lambda = \text{diag}(\lambda_0, \lambda_1, \dots, \lambda_{N-1})$ is a diagonal matrix of eigenvalues. For the discrete Laplacian defined above, the eigenvalues are:

$$\lambda_k = 4 \sin^2 \left(\frac{\pi k}{n} \right), \quad k = 0, 1, \dots, n-1.$$

These eigenvalues are non-negative ($\lambda_k \geq 0$), with $\lambda_0 = 0$ corresponding to the constant mode. Consequently, $I_N + \bar{L}_N$ is decomposed as:

$$I_N + \bar{L}_N = F(I_N + \Lambda)F^*,$$

and the fractional power of the operator is then:

$$(I_N + L_N)^s = F(I_N + \Lambda)^s F^*.$$

Thus, the norm can be computed as:

$$\|r_N\|_{H^s}^2 \approx r_N^\top (I_N + L_N)^s r_N = r_N^\top F(I_N + \Lambda)^s F^* r_N = \sum_{k=0}^{n-1} (1 + \lambda_k)^s |\hat{r}_{n,k}|^2,$$

where $\hat{r}_n = F^* r_N$ is the discrete Fourier transform of r_N . Since the fast Fourier transform (FFT) can compute $\hat{r}_n$ in $O(N \log N)$ time and other operations are $O(N)$, the overall complexity of computing the Sobolev norm is $O(N \log N)$, improving efficiency compared to direct matrix-vector multiplication, which would be $O(N^2)$.

B.4.3 Discrete Laplacian for Non-Uniform Points

In many practical applications, such as time series analysis with irregularly sampled data, the points $t_1 < t_2 < \dots < t_N$ are not uniformly spaced. In this case, we have to adjust the discrete Laplacian to account for the non-uniform spacing between points.

For a non-uniform grid, let us define the forward and backward step sizes at each point:

$$\begin{aligned} h_i^+ &= t_{i+1} - t_i, \quad i = 1, \dots, N-1, \\ h_i^- &= t_i - t_{i-1}, \quad i = 2, \dots, N, \end{aligned}$$

and define $h_1^- = h_1^+$ and $h_N^+ = h_N^-$. We introduce the coefficients

$$a_i = -\frac{2}{h_i^-(h_i^+ + h_i^-)}, \quad c_i = -\frac{2}{h_i^+(h_i^+ + h_i^-)}, \quad b_i = \frac{2}{h_i^+ h_i^-}, \quad i = 2, \dots, N-1.$$

The entries of the discrete Laplacian $L_N \in \mathbb{R}^{N \times N}$ are given by

$$L_N = \begin{bmatrix} b_1 & c_1 & 0 & \dots & 0 & 0 \\ a_2 & b_2 & c_2 & \dots & 0 & 0 \\ 0 & 0 & \dots & a_{N-1} & b_{N-1} & c_{N-1} \\ 0 & 0 & \dots & 0 & a_N & b_N \end{bmatrix}.$$

With the adjusted discrete Laplacian, we can compute the Sobolev norm in a similar manner as before.

C Experiment Details

C.1 Dataset Details

We choose several representative datasets as benchmarks, ranging from easy to difficult. These datasets include different properties like sampling rate, allowing for a more comprehensive evaluation of various semi-supervised time series classification methods. Table 6 summarizes the statistics of various datasets. The details of other 30 UCR/UEA datasets can be found in [16].

- **CricketX** [32]. The dataset contains gesture data with position of the X axis collected from accelerometers in 3D space. CricketX includes 12 classes: “Cancel Call”, “Dead Ball”, “Four”, “Last Hour Leg Bye”, “No Ball”, “One Short”, “Out”, “Penalty Runs”, “Six”, “TV Replay”. Both the training and test sets contain 390 samples.
- **UWaveGestureLibraryAll (UWave)** [30]. The dataset from the gesture recognition system used by Nokia’s search engine collects user-phone interaction motions. It contains 4,478 samples. The training and test sets contain 896 and 3,582 samples, respectively.
- **InsectWingbeatSound (InsectWing)** [12]. The dataset is released by the Computational Entomology group at the University of California Riverside for insect classification. It includes wingbeat audio signals from male and female mosquitoes, different species of flies, and other insects. The training set contains 220 samples and the test set contains 1,980 samples.
- **SelfRegulationSCP2 (SelfReg)** [7]. The University of Tuebingen releases SelfRegulation-SCP2 [7], which contains EEG data with seven columns and 1,152 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.
- **NATOPS** [20]. The AALTD competition releases the NATOPS dataset [20]. These sensors collect data from 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.
- **Heartbeat** [29]. The Heartbeat dataset 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). The sensors are sequentially positioned at the aortic, pulmonic, tricuspid and mitral auscultation sites in patients spanning a broad age range.

Table 6: The statistics of univariate and multivariate datasets in UEA & UCR archive, including three univariate datasets and three multivariate datasets for evaluation.

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

C.2 Hyperparameter Setting

We use stochastic gradient descent with a learning rate of 10^{-3} . The batch size is set to 64 with a maximum of 300 epochs. Due to the model-agnostic properties of f-VAT, we use an eight-layers Temporal Convolutional Network (TCN) [5] as the backbone architecture to compare with other competitive baselines. We run our experiments on eight NVIDIA A10 GPUs (each with 24 GB memory).

D Ablation Study

D.1 Various Architectures

To verify the model-agnostic property of our proposed f-VAT, we select several representative deep architectures with comparable parameter size, including Gated Recurrent Units (GRU), Self-Attention encoder and Temporal Convolutional Network (TCN).

- **Gated Recurrent Unit (GRU)** integrates the output and memory gates to address short-term memory challenges [15]. GRUs are combined with attention mechanisms to better capture trend information within time series data.
- **Self-Attention encoder (SA)** [39]. SA encoder is composed of four self-attention layers and a positional encoding layer, which can effectively extract trend information from modeling time series data.
- **Temporal Convolutional Network (TCN)** is a widely used for sequence modeling. Its causal convolutional operations effectively capture both short-term fluctuations and long-term temporal dependencies [5]. TCN achieves superior performance in various sequence modeling tasks.

We report the mean performance of each architecture over five runs with different random seeds. As shown in Fig. 5, TCN significantly outperforms other deep architectures on NATOPS with various label ratios. These observations show that causal convolution of TCN can effectively capture trend information, thus effectively utilizing unlabeled data to further improve generalization. In this paper, we adopt the 8-layer TCN as our main architecture.

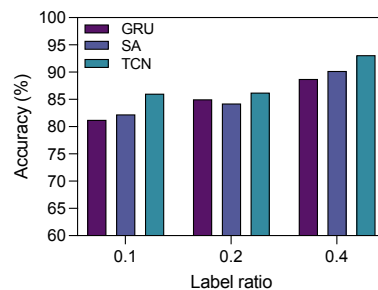


Figure 5: The performance comparison of different architectures on NATOPS with different label ratios. TCN significantly outperforms other architectures in all settings.

Table 7: Comparison of VAT, VAT-step, and f-VAT on five datasets under different label ratios.

Methods	CricketX			UWave			InsectWing			SelfReg			NATOPS		
	10%	20%	40%	10%	20%	40%	10%	20%	40%	10%	20%	40%	10%	20%	40%
VAT	42.85 ± 3.07	49.14 ± 0.50	58.63 ± 0.50	94.41 ± 0.09	95.53 ± 0.31	94.76 ± 0.54	55.49 ± 1.28	61.27 ± 0.19	63.48 ± 0.30	53.12 ± 4.51	55.76 ± 0.35	53.47 ± 1.04	82.38 ± 0.06	82.81 ± 0.52	90.15 ± 1.00
VAT-step	40.47 ± 3.84	47.85 ± 0.33	59.30 ± 0.36	93.78 ± 0.07	96.09 ± 0.32	95.86 ± 0.55	57.45 ± 0.21	60.92 ± 0.48	63.61 ± 0.58	48.61 ± 3.27	53.67 ± 2.79	52.75 ± 1.80	81.33 ± 0.06	83.02 ± 0.12	86.58 ± 1.18
f-VAT	49.18 ± 1.96	57.91 ± 3.58	68.39 ± 2.25	94.82 ± 0.30	96.45 ± 0.27	97.23 ± 0.43	58.01 ± 1.12	61.28 ± 1.86	64.81 ± 1.15	59.31 ± 3.06	61.60 ± 1.13	64.44 ± 3.13	86.04 ± 1.41	86.25 ± 1.38	93.13 ± 0.15

To verify the straightforward improvement to VAT may be unsuitable for time series, we design a variant called *VAT-step*, which constrains the magnitude of perturbations at each time step, and conduct extensive experiments across multiple datasets with different label ratios. Table 7 shows that VAT-step only provides marginal performance gains compared to the original VAT in most settings, and even degrades performance on more challenging multivariate datasets (e.g., SelfReg and NATOPS) containing complex temporal structure. This is because clipping the perturbation at each time step struggles to adaptively scale the magnitude on critical regions for prediction (like “shapelets” in Fig. 4), unless the per-step magnitude hyperparameters are carefully tuned. Additionally, VAT-step can easily converge to the “permutation-invariant” [8] local optimum, where arbitrarily reordering time steps can still generate identical perturbations satisfying each step-wise constraint. By contrast, adversarial perturbations generated by *f-VAT*, which incorporate the temporal structure of time series, more efficiently utilize sequential information to improve the smoothness of predictive distributions and the final predictive performance.

that the extra computational cost remains in the same order as VAT. The proposed f-VAT achieves competitive performance without incurring significant computational costs, making it particularly suitable for semi-supervised time series classification with limited computational resources.

Table 11: Runtime comparison of f-VAT and VAT.

Method	CricketX	UWave	InsectWing	NATOPS	SelfReg
VAT	15.67	51.66	35.58	28.04	30.68
f-VAT	20.45	62.05	45.92	38.98	44.95
Δ (%)	30.50	20.11	29.06	39.02	46.51