Long-tailed Time Series Classification via Feature Space Rebalancing

Pengkun Wang[1], Xu Wang[1], Binwu Wang[1], Yudong Zhang[1], Lei Bai[2(✉)], and Yang Wang[1(✉)]

[1] University of Science and Technology of China, Hefei 230000, China
{pengkun, wx309, wbw1995, zyd2020}@mail.ustc.edu.cn, angyan@ustc.edu.cn
[2] Shanghai AI Laboratory, Shanghai 200000, China
baisanshi@gmail.com

Abstract. Learning unbiased decision boundaries is crucial for time series classification. Real-world datasets typically exhibit long-tailed natures of class distributions, which results in an imbalanced feature space after training, i.e., decision boundaries will be easily biased towards dominant classes that dominate the feature space. However, existing methods mostly train models from artificially balanced datasets, making it still unclear how to deal with the long-tailed natures of time series data in real-world scenarios. Motivated by this question, we analyze the similarities and differences between long-tailed time series classification and general long-tailed recognition, and propose a Feature Space Rebalancing (FSR) strategy for time series classification, which works jointly from both representation and data perspectives. Specifically, from the representation perspective, we design Balanced Contrastive Learning (BCL), which avoids excessive intra-class compaction of tail classes by introducing a balanced supervised contrastive loss with hierarchical prototypes, resulting in a balanced feature space and better generalization. From the data perspective, we explore the effectiveness of traditional data augmentation on long-tailed distributions and propose an Adaptive Temporal Augmentation (ATA) to rebalance the potential feature space at the temporal level. Extensive experiments on multiple long-tailed time series datasets demonstrate its superiority, including different class distributions and imbalance ratios.

Keywords: Time series classification · Long-tailed recognition · Contrastive learning.

1 Introduction

Time series classification (TSC) has been widely explored as it is associated with massive real-world applications and has a significant impact on human life [4, 8, 17, 1, 18, 26, 28, 9, 12]. For instance, some recent TSC methods [4, 26] have spared no effort to introduce deep neural networks into the medical field, and realize

Yang Wang is the corresponding author. Lei Bai is the joint corresponding author.

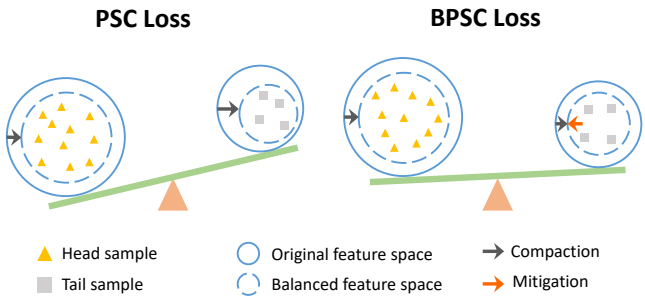

Fig. 1. BPSC loss dynamically rebalances the feature space with a class-dependent compaction factor. It is 'forgiving' to tail classes and avoids the final feature spaces of tail classes becoming over-compacted and biased.

intelligent disease diagnosis by automatically classifying the ECG signals of patients. Meanwhile, massive real data make it possible to train more complex deep models. Therefore, researchers are focusing on designing exquisite network structures or extracting semantically precise high-dimensional representations to further improve the discriminativeness of models.

Despite significant progress, most existing TSC methods [8, 26] focus on learning unbiased decision boundaries from artificially balanced datasets (i.e. all the classes have similar sample sizes). However, in the real world, class distributions of time series data typically exhibit long-tailed nature, which makes the decision boundaries easily biased towards the dominant classes with massive training data and thus decreases the classification accuracy. Although a few methods have considered class imbalance, they either only explored simple tasks with two categories [12, 9] or seriously ignore the long-tailed nature [30]. Thus, several issues are worth pondering, e.g., 1) *What are the challenges of long-tailed time series classification relative to the other domain (especially vision)*? 2) *Are general long-tailed recognition methods applicable to the time series domain*? 3) *How to realize efficient long-tailed time series classification*? These issues are closely related to the scalability and generalization ability of long-tailed learning but have not been well-explored in the TSC area.

Motivated by the above problems, we analyze the **similarities** and **differences** between long-tailed time series classification (Long-tailed TSC) and general long-tailed recognition (GLR). When ignoring the difference in the data dimension, Long-tailed TSC and GLR can be regarded as homogeneous problems. For example, we can extract label-dependent information based on the overall class distribution of the dataset to achieve supervised inter-class balance or apply the idea of clustering to realize weakly-supervised/unsupervised inter-class balance. Therefore, existing methods of resampling or improving general cost-sensitive functions have certain generalizations to Long-tailed TSC, e.g., the improved Softmax for multi-class probability calculation [22]. However, unlike images, time series have unique temporal properties, which determines that we cannot solve Long-tailed TSC from a sample perspective alone. From the per-

spective of the data dimension, existing methods alleviate the long-tailed nature by designing complex network structures or semantically related processing modules, but they cannot be directly transferred to time series due to the inability to model temporal information. Moreover, the correlation information between variables is also impossible to model by existing long-tailed methods. The above inspires us to consider this peculiar property when solving Long-tailed TSC.

Motivated by this, based on a comprehensive consideration of temporal properties, we propose a novel Feature Space Rebalancing (FSR) strategy for long-tailed time series classification that works jointly from both representation and data perspectives. From the representation perspective, we design a Balanced Contrastive Learning (BCL) that consists of two key parts: balanced prototypical supervised contrastive (BPSC) loss and hierarchical prototypes. The former adjusts the degree of intra-class compaction by a class-dependent compaction factor when computing the intra-class prototype similarity, thus avoiding an imbalanced feature space. The latter considers the unique temporal properties of time series and introduces additional temporal prototypes when computing the contrastive loss, which are extracted by a simple temporal module. These hierarchical prototypes characterize time series more comprehensively, bringing the learned feature space closer to the essence of time series. These hierarchical prototypes comprehensively characterize time series, bringing the learned feature space closer to the essence of time series.

We also rebalance the feature space from the data perspective. It is well known that traditional data augmentation (e.g. jittering) can expand the feature space and improve the intra-class diversity, thereby making the class distribution closer to the true distribution [25]. It is a common practice to apply the same degree of data augmentation to all classes. However, there is a fact that is easily neglected in long-tailed datasets, that is, the feature distribution of the head class is relatively close to the true distribution because of the massive samples, while the feature distribution of the tail class may be biased because of the sparse samples and poor intra-class diversity. We propose Adaptive Temporal Augmentation (ATA) to alleviate this problem. The core idea is to assign different degrees of temporal augmentation (e.g. jittering) to each class according to the sample size, thus improving the intra-class diversity of the tail class and balancing the augmented feature space. Specifically, for a multivariate time series, we set independent degree-consistent augmentation for each variable, and the degree is determined by its label information.

Our contributions are summarized as follows:

- We comprehensively discuss the long-tailed time series classification learning and construct three corresponding long-tailed datasets. To the best of our knowledge, this is the first long-tailed time series classification work, which fills a gap in the field.
- To address the above Long-tailed TSC, we propose a novel Feature Space Rebalancing (FSR) strategy. First, we design a Balanced Contrastive Learning (BCL) from the representation perspective, which avoids imbalanced feature spaces by introducing compaction factors and hierarchical prototypes

in the supervised contrastive loss. Second, from the data perspective, we rethink traditional data augmentation and propose an Adaptive Temporal Augmentation (ATA) to balance the augmented feature space.
– We conduct extensive experiments on the three proposed datasets and demonstrate that the proposed FSR is more suitable for long-tailed time series classification than existing methods.

2 Related Works

Time Series Classification. In recent years, extensive studies have been made on time series classification with deep neural networks [8, 17, 1, 18, 26]. These methods aim to achieve better model performance by designing exquisite network structures [1, 18, 26] or improving plug-and-play modules [8, 17]. However, state-of-the-art methods mainly experiment with balanced datasets to demonstrate their capacity, ignoring the imbalance problem of real-world datasets. To this end, several studies propose diverse strategies to address imbalanced time series classification [9, 12, 30]. But these methods mainly focus on scenarios with a small number of classes (e.g. 2), while real-world scenarios usually have a large number of classes. In comparison, we construct three time series datasets with different class distributions and imbalance ratios for long-tailed natures, and propose FSR to model complex data distributions.

Long-tailed Recognition. In real-world scenarios, class distributions typically exhibit long-tailed natures, which makes the trained model easily biased toward head classes with massive data [29]. Many methods have made efforts to address this class imbalance and they can be grouped into three categories: class re-balancing [20, 6, 3, 22], information augmentation [27], and module improvement [24, 14, 19]. In the field of class re-balancing, a mainstream strategy is to design cost-sensitive loss functions to adjust loss values for different classes during training, e.g., CB loss [6] or balanced softmax loss [22]. Further, some methods combine improved cost-sensitive loss functions with well-designed network structures to achieve efficient long-tailed learning, e.g., decoupled training [15] or ensemble learning [24]. Most recent studies focus on general long-tailed recognition, however, limited effort has been made for long-tailed time series classification due to the lack of proper benchmarks. Inspired by this, we construct three benchmarks to fill the gap and design hierarchical prototypes based on temporal properties to ensure semantic consistency in contrastive learning.

Contrastive Learning. Contrastive learning has achieved outstanding success in self-supervised representation learning, which has profound implications for a variety of downstream tasks [5]. The basic principle of contrastive learning is to learn a high-dimensional semantic feature space by constructing positive and negative sample pairs, and attract the positive sample pairs and repulsing the negative sample pairs. Recent works also found that using contrastive loss

in long-tailed learning can obtain representation models generating a better feature space [14, 24, 19, 13]. It is worth noting that Hybrid [24] proposes a hybrid network structure with a prototypical supervised contrastive loss, which resolves the memory bottleneck resulting from standard supervised contrastive learning. However, this loss is not friendly to feature space balancedness, since imposing the same degree of intra-class constraints on all the classes would result in excessive intra-class compaction of tail classes. In our work, we focus on the adaptability in representing distance computations and propose a balanced prototypical supervised contrastive loss to avoid excessive intra-class compaction of tail classes.

3 Long-tailed Time Series Classification

Conventional time series classification cannot cope with the long-tailed natures in real-world applications, resulting in poor performance of the trained model on tail classes. However, tail classes are critical for tasks such as abnormal activities in behavior recognition and rare conditions in disease diagnosis. Considering that no research has been explicitly investigated in this direction so far, we give a detailed problem definition and corresponding datasets to fill this gap.

3.1 Problem Definition

Conventional time series classification methods mostly train models on balanced datasets. Differently, long-tailed time series classification focuses on training a robust deep neural network from a time series dataset with a long-tailed class distribution. This long-tailed nature can be understood as the fact that a small number of classes have massive time series samples while other classes are only related to a few samples. More formally, let $\{x_i, y_i\}_{i=1}^{N}$ be the long-tailed time series training set, where each time series x_i corresponds to a class label y_i. Assuming that a dataset contains $\mathcal{C}$ classes, the sample number of class c is n_c, and the total number of the entire dataset is $N = \sum_{k=1}^{\mathcal{C}} n_c$, the imbalance ratio of the time series dataset can be defined as n_{max}/n_{min}, where n_{max} and n_{min} denote the sample size of the class containing the most and least samples, respectively. Without loss of generality, the class distribution in the training set exhibits a long-tailed trend when sorted by cardinality in decreasing order. Additionally, a time series dataset may be univariate or multivariate.

3.2 Proposed Datasets

Based on existing datasets, we construct three derived long-tailed time series classification datasets to fill the gaps in this field. Referring to mainstream datasets, we divide the classes of each dataset into head classes, medium classes, and tail classes according to the sample size. The visualized class distributions and statistics are shown in Figure 2 and Table 1 respectively.

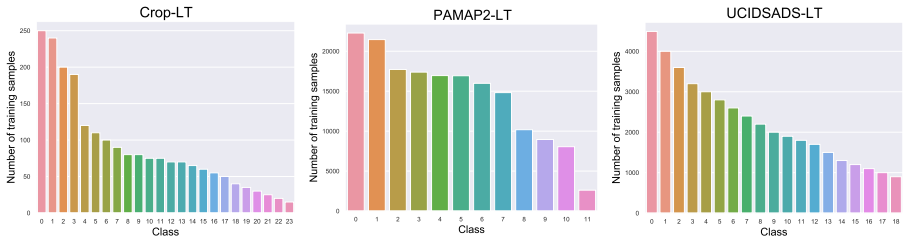

Fig. 2. Class distributions of proposed long-tailed time series classification datasets.

Table 1. Statistics of proposed datasets ($\star$ and # denote the "number" and "sample size" respectively).

Dataset	Crop-LT	PAMAP2-LT	UCIDSADS-LT
Variable$\star$	1	36	45
Class$\star$	24	12	19
Length	46	20	20
Training set$\star$	2,145	173,309	42,692
Validation set$\star$	1,200	6,000	9,500
Test set$\star$	16,800	12,000	19,000
Head classes$\star$	2 (# >200)	2 (# >20000)	4 (# >3000)
Medium classes$\star$	5 (200$\geq$ # $\geq$100)	6 (20000$\geq$ # $\geq$11000)	9 (3000$\geq$ # $\geq$1700)
Tail classes$\star$	17 (# <100)	4 (# <11000)	6 (# <1700)
Imbalance ratio	17	9	5

Crop-LT. Crop is a univariate dataset from the well-known UCR time series archive [7]. It consists of 24 classes with the same sample size and the length of each sample is 46. To better evaluate long-tailed time series classification methods, we resample a long-tailed training set, i.e., Crop-LT, which has 2 head classes, 7 medium classes, and 17 tail classes. The imbalance ratio is 17. The overall training set size is 2,145.

PAMAP2-LT. PAMAP2 is a multivariate benchmark for daily physical activity classification, and its data is collected with three IMUs placed on the subject's chest, dominant wrist, and dominant ankle, respectively, under the sampling frequency of 100Hz [21]. The sampled dataset PAMAP2-LT contains 12 classes with a total of 173,309 training data and an imbalance ratio of 9. In PAMAP2-LT, each time series are acquired by 36 sensors, and the time step length of each stream is 20. Similar to the Crop-LT dataset, we define 2 head classes, 6 medium classes, and 4 tail classes.

UCIDSADS-LT. UCIDSADS is a multivariate benchmark specially devised for daily and sports activities, which comprises the motion sensor data of 19 daily and sports activities [2]. The samples in this benchmark are acquired by 45 sensors at the sampling frequency of 25Hz. We sample 42,692 time series from

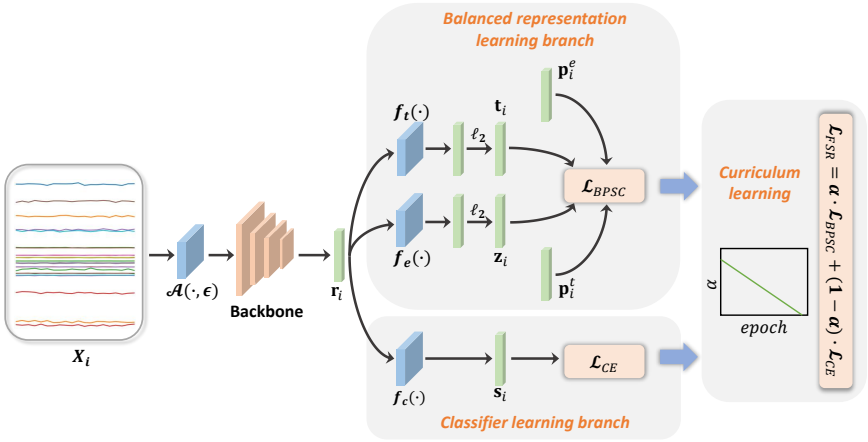

Fig. 3. Overview of the proposed feature space rebalancing based network structure. The network is hybrid, and it contains a balanced representation learning branch and a classifier learning branch. The former utilizes balanced contrastive learning to learn a balanced and unbiased feature space, while the latter is a traditional classification strategy. During training, the curriculum learning is used to achieve a smooth transition from the balanced representation learning branch to the classifier learning branch.

the original dataset to form the UCIDSADS-LT dataset with an imbalance ratio of 5, and the length of each time series is 20. Like the PAMAP2-LT dataset, we define 4 head classes, 9 medium classes, and 6 tail classes.

4 Methodology

Based on the above problem definition, we try to solve the problem: *How to realize efficient long-tailed time series classification*? To this end, we propose a feature space rebalancing (FSR) strategy that consists of two parts: balanced contrastive learning (BCL) and adaptive temporal augmentation (ATA). Motivated by [24], we design an improved hybrid network for Long-tailed TSC as shown in Figure 3. Furthermore, in the initialization phase, we incorporate the ATA module into the framework.

4.1 Balanced Contrastive Learning

Balanced contrastive learning aims to learn a balanced feature space that balances head and tail classes while achieving intra-class compactness and inter-class separability, thereby helping the classifier learn unbiased decision boundaries. For a time series (x_i, y_i), its corresponding high-dimensional representation $\mathbf{r}_i$ can be generated by various backbone networks, such as the widely used Temporal Convolutional Network (TCN) [16], LSTM [11], and more powerful and

advanced Transformer models [32]. Since designing more powerful representation networks is parallel with long-tailed learning, we simply utilize a shared backbone network (e.g. ResNet-TS [28]) to learn $\mathbf{r}_i$ considering its stable representation ability and fair comparison with existing methods, which drives the learning of the balanced time series representation learning branch and classifier learning branch as shown in Figure 3.

Hierarchical Prototypes. For the balanced time series representation learning branch, a nonlinear multiple-layer perceptron (MLP) $f_e(\cdot)$ combined with ℓ_2 normalization is regarded as a projection head to map $\mathbf{r}_i$ into a vector representation $\mathbf{z}_i$, which is more suitable for contrastive learning. Considering the unique temporal properties of time series, we additionally use an LSTM $f_t(\cdot)$ to extract the corresponding temporal representation $\mathbf{t}_i$ from $\mathbf{r}_i$. $f_t(\cdot)$ aggregates high-dimensional representations into compact representations, which also combines with ℓ_2 normalization.

$$\mathbf{z}_i = \ell_2(f_e(\mathbf{r}_i)), \quad \mathbf{t}_i = \ell_2(f_t(\mathbf{r}_i)). \tag{1}$$

For these two different perspectives of representation, we utilize their average representation as prototype representations, which are defined as $\mathbf{p}_i^e$ and $\mathbf{p}_i^t$.

Balanced Prototypical Supervised Contrastive Loss. As we mentioned, the prototypical supervised contrastive (PSC) loss can resolve the memory bottleneck issue by learning a prototype for each class [24]. For a long-tailed dataset with $\mathcal{C}$ classes, the goal of PSC is to learn a prototype feature for each class during training and guide the vector representation to be closer to the prototype of their class and far away from the prototypes of other classes. The formulation of PSC loss can be written as

$$\mathcal{L}_{PSC}(\mathbf{z}_i) = -log\frac{e^{(\mathbf{z}_i \cdot \mathbf{p}_{y_i}/\tau)}}{\sum_{j=1,j\neq y_i}^{\mathcal{C}} e^{(\mathbf{z}_i \cdot \mathbf{p}_j/\tau)}}, \tag{2}$$

where $\tau > 0$ is a scalar temperature parameter, and $\mathbf{p}_{y_i}$ is the prototype representation for class y_i, which is normalized to the unit hypersphere.

Although reducing memory consumption, PSC imposes the same degree of intra-class constraints on all the classes. As shown in Figure 1, for the head classes, this constraint enforces a more compact feature space, thus mitigating their intrusion into tail classes. For the tail classes, their feature spaces also become more compact. However, the prototype representations of these tail classes learned by the model may be biased due to the small sample size, which leads to the final feature spaces of these classes being over-compacted and biased. Overall, this balanced constraint will exacerbate the feature space imbalancedness and impair the generalization of the model on the tail classes.

To alleviate the above-mentioned problem, we propose to impose different degrees of constraints on different classes and design a balanced prototypical

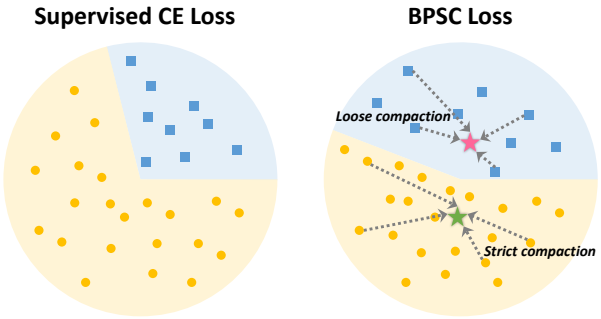

Fig. 4. Feature spaces learnd by our BPSC loss. Compared with the supervised cross-entropy (CE), our BPSC loss learns a balanced feature space. The *star* indicates the prototype of each class, and the *shadow area* indicates the decision boundary.

supervised contrastive (BPSC) loss as

$$\begin{aligned}\mathcal{L}_{BPSC}(\mathbf{z}_i) = &- log\frac{\omega_i \cdot e^{(\mathbf{z}_i \cdot \mathbf{p}^e_{y^i}/\tau)}}{\sum_{j=1,j\neq y_i}^{\mathcal{C}} e^{(\mathbf{z}_i \cdot \mathbf{p}^e_j/\tau)}} \\ &+ \beta \cdot (-log\frac{\omega_i \cdot e^{(\mathbf{z}_i \cdot \mathbf{p}^t_{y_i}/\tau)}}{\sum_{j=1,j\neq y_i}^{\mathcal{C}} e^{(\mathbf{z}_i \cdot \mathbf{p}^t_j/\tau)}}) \\ with \quad \omega_i &= \frac{n_{max}}{n_i},\end{aligned} \tag{3}$$

where β is a weighting coefficient of the temporal property, and ω_i represents the compaction factor, which leads the contrastive loss to be more 'forgiving', i.e., maintaining strict intra-class compaction for head classes and relatively loose intra-class compaction for tail classes. According to the theory that a balanced feature space helps to learn high-quality representations, the compaction factor can avoid excessive intra-class compaction of tail classes to a certain extent, resulting in a balanced inter-class feature space, as illustrated in Figure 4.

Nevertheless, we found that the compaction factor of the tail class tends to be 0 when the imbalance rate of the dataset is high, which leads to overly loose feature space and thus fails to learn high-quality representations. For this mixed blessing, we propose a mitigative compaction factor

$$\omega_i = e^{-n_{max}/n_i \cdot \rho_b}. \tag{4}$$

By adjusting the mitigation coefficient ρ_b, we can obtain the optimal intra-class compaction which is beneficial to balancing the feature space and learning high-quality representations.

Moreover, the classifier learning branch is simpler which applies a single linear layer $f_c(\cdot)$ to $\mathbf{z}_i$ to predict the class-wise logits $\mathbf{s}_i$. During the training process, we also employ a curriculum [31] to adjust the weightings of these two branches to realize a smooth transition from balanced representation learning to classifier

learning. The final loss function is jointly determined by the two branches:

$$\mathcal{L}_{FSR} = \alpha \cdot \mathcal{L}_{BPSC} + (1 - \alpha) \cdot \mathcal{L}_{CE}, \tag{5}$$

where α is a weighting coefficient inversely proportional to the number of epochs.

4.2 Adaptive Temporal Augmentation

Mainstream long-tailed learning methods focus on model structure and representation, while data augmentation has received little attention. It has been proven that traditional data augmentation can enlarge the feature space and increase the intra-class diversity, and is beneficial to improving the generalization of the model [25]. A consensus approach is to apply the same degree of data augmentation to all the classes, which is not suitable for long-tailed datasets due to high imbalance ratios. Specifically, the feature space of the tail class is usually smaller than that of the head class for the long-tailed dataset. Adding the same degree of augmentation will expand the feature space of both the head and tail classes, leaving the feature space imbalancedness still existing. If we want to get a balanced feature space, it is necessary to balance the inter-class diversity as much as possible, which is similar to the idea of resampling.

To achieve this goal, we propose an Adaptive Temporal Augmentation (ATA) to assign different degrees of augmentation to each class according to the sample size. We define a parametric temporal augmentation method (e.g. jittering) as $\mathcal{A}(\cdot, \epsilon)$, where ϵ is the augmentation factor. Then, the augmented sample is

$$\hat{x_i} = \mathcal{A}(x_i, \epsilon). \tag{6}$$

For example, when using temporal jittering augmentation, we append an independent degree-consistent noise sequence to each variable in a multivariate time series.

In traditional data augmentation, the augmentation factor ϵ is a constant for all classes. In adaptive temporal augmentation, our goal is to balance the inter-class diversity, so the augmentation factor of the head class will be smaller than that of the tail class, thus maintaining the stability of the head class and the diversity of the tail class. Similar to BCL, it is better to make the model have better generalization performance on the head class, so we propose a mitigative augmentation factor

$$\epsilon_i' = \epsilon \cdot e^{-\frac{n_i}{n_{max}} \cdot \rho_a}, \tag{7}$$

where ρ_a is the mitigation factor. If $n_i > n_j$, then $\epsilon_i' < \epsilon_j'$. The mitigative augmentation factor forces the inter-class diversity to be closer, and ultimately, the trained model not only learns a more balanced feature space but also recognizes tail classes generalized.

5 Experiments

Implementation details. For all three datasets, for a fair comparison with existing long-tailed learning methods, we use the general time series feature

Table 2. Performance on univariate Crop-LT dataset and multivariate PAMAP2-LT and UCIDSADS-LT datasets.

Method	Crop-LT				PAMAP2-LT				UCIDSADS-LT			
	Head	Medium	Tail	All	Head	Medium	Tail	All	Head	Medium	Tail	All
CE [10]	89.65	46.77	57.93	58.25	89.25	71.30	74.25	75.28	82.15	72.87	70.51	74.08
Focal [20]	90.15	46.37	53.42	55.01	93.90	72.90	73.83	76.71	85.45	70.06	65.45	71.84
CB [6]	79.43	38.63	63.19	59.43	90.75	70.87	76.73	76.13	79.85	72.67	72.77	74.21
LDAM [3]	85.93	45.11	62.14	60.58	85.65	71.17	73.38	74.32	77.95	71.23	66.82	71.25
BS [22]	81.14	44.60	63.71	61.18	89.35	69.80	78.30	75.89	78.08	76.28	67.35	73.84
Seesaw [23]	85.29	47.69	61.53	60.62	92.35	72.12	75.03	76.46	79.70	72.14	80.53	76.38
Hybrid [24]	87.57	42.14	59.96	58.55	92.35	69.40	73.40	74.56	80.22	72.04	67.53	72.34
KCL [14]	88.21	47.98	62.84	61.85	92.15	70.82	75.27	75.86	80.33	71.57	71.95	73.53
TSC [19]	88.01	48.10	63.27	62.17	91.80	71.03	77.24	76.56	81.59	74.37	72.79	75.39
FSR	89.75	50.29	66.31	**64.93**	93.10	73.70	80.55	**79.05**	82.55	77.48	78.57	**78.89**

extraction network ResNet-TS [28] as the backbone network. For FSR, $f_e(\cdot)$ is a nonlinear MLP with one hidden layer, $f_t(\cdot)$ is a two-layer LSTM, and $f_c(\cdot)$ is a single linear layer. We use Pytorch to implement all neural networks and train the model on 8 NVIDIA Tesla V100 GPUs. The networks are trained for 200 epochs by the Adam optimizer with a learning rate of 10^{-4} and weight decay of 4×10^{-3}. For Crop-LT dataset, the batch size is 128, the weighting coefficient α is $1 - (Epoch_{now}/Epoch_{max})$, and β is 0.5. For PAMAP2-LT and UCIDSADS-LT datasets, the batch size is 256, the weighting coefficient α is $1 - (Epoch_{now}/Epoch_{max})^2$, and β is 0.9. In BCL, we set the mitigation coefficient ρ_b to be 0.5. In ATA, we use a jittering with an augmentation factor ϵ of $(0, 0.1)$ and mitigation factor ρ_a of 1 as augmentation.

Compare with state-of-the-art methods. The comparison between the proposed FSR and state-of-the-art methods on three long-tailed datasets is presented in Table 2. Based on the partitioning of the datasets above, we show the average accuracy on all classes, and also on each subset. For a comprehensive comparison, in addition to the cross-entropy (CE) loss, we select a variety of long-tailed recognition methods as baselines, which are based on *different theoretical ideas*, including class-level re-weighting [20, 6, 23], class-level re-margining [3],class-balanced re-sampling [22], metric learning [24], and decoupled training [14, 19]. And our proposed FSR can be classified into metric learning or a new class-balanced augmentation.

As can be seen from the table, on the univariate Crop-LT dataset, the existing long-tailed learning methods can generally improve the overall classification accuracy compared to CE. FSR is no exception, it outperforms the compared methods on the medium and tail subsets. In particular, it outperforms CE by 3.52% and 8.38%, respectively. However, the performance of the compared methods is unstable on the multivariate PAMAP2-LT and UCIDSADS-LT datasets. It

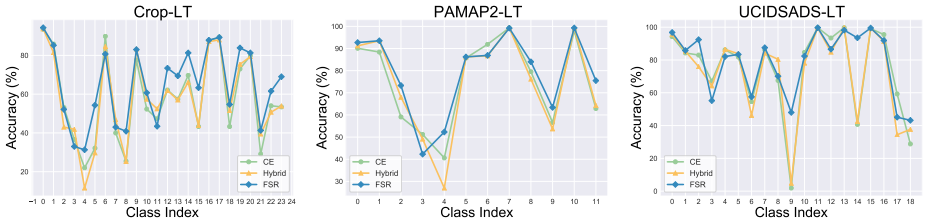

Fig. 5. Visualization of the accuracy of each class on three proposed long-tailed time series classification datasets.

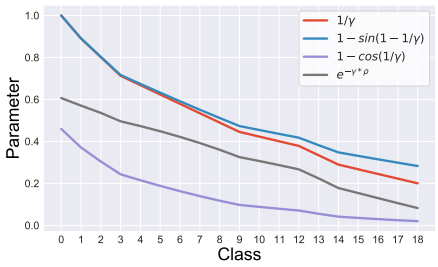

Fig. 6. Visualization trends of the proposed mitigation compaction factor and its variants.

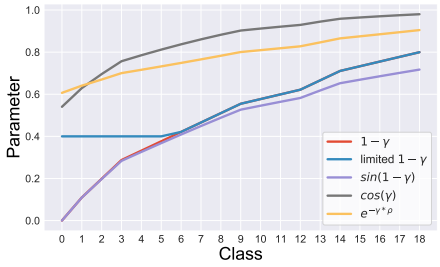

Fig. 7. Visualization trends of the proposed mitigative augmentation factor and its variants.

is obvious that the method based on the improved loss function cannot effectively improve the performance on the tail class, due to the imbalanced feature space learned by the single-stage classifier learning. The recent mainstream methods based on contrastive learning have a steady improvement on the tail class, but the overall accuracy is not excellent. The reason is that although a balanced feature space is learned, the temporal information is ignored. Our methods address such limitations in that: 1) Hierarchical prototypes that consider temporal information; 2) Class-dependent intra-class compaction that balances the feature space; 3) Adaptive augmentation that improves tail class diversity.

Visualization of accuracy on each class. To more intuitively observe the superiority of FSR, we visualize the accuracy of each class. It can be found that compared with baselines, the accuracy of FSR is more stable, which means that it balances all classes as much as possible, especially the difficult ones. In the head class, the accuracy of FSR is slightly better than Hybrid, because the mitigative augmentation factor also improves the diversity of the head class. In the medium and tail classes, FSR significantly outperforms baselines, which benefits from the 'forgiving' of balanced contrastive learning for these classes.

Mitigative compaction factor for BCL To illustrate the rationality of the proposed mitigative compaction factor in BCL, we compare different variants

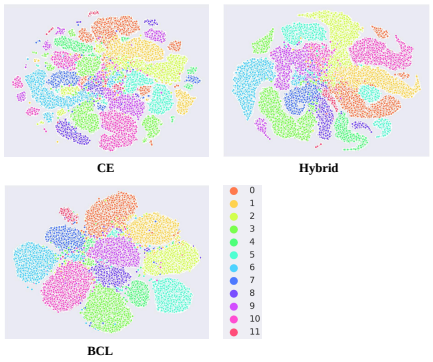

Fig. 8. T-SNE visualizations of representations of CE, Hybrid, and our proposed BCL on the PAMAP2-LT dataset. Different colors represent different classes.

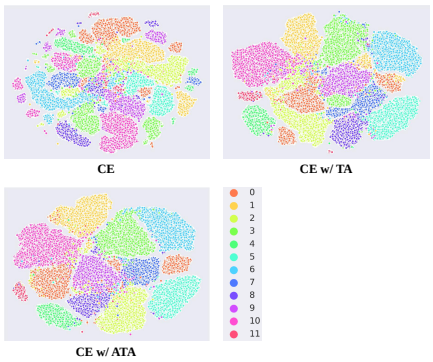

Fig. 9. T-SNE visualizations of representations of CE, CE with TA, and CE with ATA on the PAMAP2-LT dataset. Different colors represent different classes.

Table 3. Comparison of the proposed mitigation compaction factor and its variants on the UCIDSADS-LT dataset.

δ	Head	Medium	Tail	All
Hybrid	80.22	72.04	67.53	72.34
$1/\gamma$	80.17	74.39	72.82	75.11
$1 - sin(1 - 1/\gamma)$	80.30	74.87	73.62	75.62
$1 - cos(1/\gamma)$	80.89	76.78	74.05	76.78
$e^{-\gamma \cdot \rho}$	81.75	77.50	76.43	78.06

Table 4. Comparison of the proposed mitigative augmentation factor and its variants on the UCIDSADS-LT dataset.

δ	Head	Medium	Tail	All
CE	82.15	72.87	70.51	74.08
$1 - \gamma$	79.50	73.43	75.33	75.30
limited $1 - \gamma$	81.14	73.60	76.21	76.01
$sin(1 - \gamma)$	80.16	73.57	75.92	75.70
$cos(\gamma)$	81.50	73.81	76.47	76.27
$e^{-\gamma \cdot \rho}$	82.55	74.00	77.30	76.84

on the UCIDSADS-LT dataset. Here, we use γ to denote n_{max}/n_i. Figure 6 shows the changing trend of different variants. From the results in Table 3, the proposed compaction factor can significantly alleviate the imbalance problem caused by traditional compaction. In addition, the mitigative compaction factor ensures that the tail classes can also be reasonably compacted.

Mitigative augmentation factor in ATA. In ATA, the mitigative augmentation factor is crucial for adaptive augmentation. Here, we use γ to denote n_i/n_{max}, and δ to denote the dynamic coefficient of ϵ_i', then

$$\epsilon_i' = \epsilon \cdot \delta = \epsilon \cdot e^{-\gamma \cdot \rho_a}. \tag{8}$$

To illustrate the rationality of the proposed dynamic coefficient, we compare different variants of δ on the UCIDSADS-LT dataset. As shown in Figure 7, these variants have different trends, some increase rapidly from 0, and some increase slowly from a lower limit. Here, we apply these variants to CE. From the results in Table 4, we have two intuitive findings: 1) the class-dependent augmentation

Table 5. Ablation results on the PAMAP2-LT dataset.

Method	Head	Medium	Tail	All
BCL	92.40	71.45	78.23	77.20
w/o BPSC & HP	92.35	69.40	73.40	74.56
w/o BPSC	92.48	70.75	76.22	76.20
w/o HP	92.27	71.24	77.87	76.96

Table 6. Comparison between baselines, baselines with TA, and baselines with ATA.

Method	Crop-LT				UCIDSADS-LT			
	Head	Medium	Tail	All	Head	Medium	Tail	All
CE	89.65	46.77	57.93	58.25	82.15	72.87	70.51	74.08
w/ TA	90.28	46.06	58.57	58.61	77.07	73.40	75.77	74.92
w/ ATA	89.72	42.08	60.97	59.43	82.55	74.00	77.30	76.84
BCL	87.86	49.29	63.96	62.90	81.75	77.50	76.43	78.06
w/ TA	86.07	45.14	64.29	62.11	81.33	75.57	74.95	76.59
w/ ATA	89.75	50.29	66.31	64.93	82.55	77.48	78.57	78.89

factor can improve the generalization of the model; 2) the fixed lower limit or mitigative increase from a lower limit is a better choice because it can increase the diversity of all classes and not just the tail class.

T-SNE Visualization. To demonstrate that the representations learned by BCL and ATA can distinguish different classes in the latent space, we visualize the representations of different methods on the PAMAP2-LT dataset by T-SNE. As shown in Figure 8, compared with Hybrid, our proposed BCL based on the compaction factor learns a more balanced feature space, so that the model can better recognize the tail classes. The effect of ATA is also significant, in Figure 9, we can observe that the feature space of the tail class is expanded without being suppressed by the head class.

Ablation study for balanced contrastive learning. In balanced contrastive learning, balanced prototypical supervised contrastive (BPSC) loss and hierarchical prototypes (HP) are the core of recognizing long-tailed time series and rebalancing feature space. Without using ATA, we compare BCL and its variants on the PAMAP2-LT dataset. As shown in Table 5, when only BPSC is used, the accuracy of the model on the medium and tail classes is significantly improved, proving that BPSC can obtain a more balanced feature space than PCS. When only HP is used, both prototypes accurately model the time series and correct the bias of the model, resulting in an overall improvement in accuracy on each subset. And when neither is used, it is obvious that the performance of the model degrades significantly.

Ablation study for adaptive temporal augmentation. To demonstrate the effectiveness of adaptive temporal augmentation (ATA), we apply ATA and traditional temporal augmentation (TA) to CE and BCL, respectively. The experimental results on Crop-LT and UCIDSADS-LT datasets are shown in Table 6. For CE, although TA improves the generalization of the model, ATA achieves a more significant improvement, especially for tail classes. This phenomenon proves that ATA helps to improve the diversity of tail classes, thereby making the feature space more balanced. For BCL, using TA leads to a decrease in the accuracy of the model, while using ATA still improves performance

steadily. From a representation perspective, we argue that a balanced and unbiased feature space helps learn accurate prototypes, while the prototypes learned by TA are biased, which leads to a significant decrease in model performance.

6 Conclusion

In this work, we construct three long-tailed time series classification datasets and propose a feature space rebalancing strategy, FSR. To the best of our knowledge, this is the first long-tailed time series classification method, which fills a gap in the field. We rebalance the feature space from two perspectives, including representation-based balanced contrastive learning and data-based adaptive temporal augmentation. Experiments on the three proposed datasets demonstrate the superiority of FSR.

Acknowledgements This paper is partially supported by the National Natural Science Foundation of China (No.62072427, No.12227901), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), Academic Leaders Cultivation Program, USTC.

References

1. Bai, L., Yao, L., Wang, X., Kanhere, S.S., Guo, B., Yu, Z.: Adversarial multi-view networks for activity recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (2020)
2. Barshan, B., Yüksek, M.C.: Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. Comput. J. (2014)
3. Cao, K., Wei, C., Gaidon, A., Aréchiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Proc. of NeurIPS (2019)
4. Chen, H., Huang, C., Huang, Q., Zhang, Q., Wang, W.: Ecgadv: Generating adversarial electrocardiogram to misguide arrhythmia classification system. In: Proc. of AAAI (2020)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: Proc. of ICML (2020)
6. Cui, Y., Jia, M., Lin, T., Song, Y., Belongie, S.J.: Class-balanced loss based on effective number of samples. In: Proc. of CVPR (2019)
7. Dau, H.A., Bagnall, A.J., Kamgar, K., Yeh, C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Keogh, E.J.: The UCR time series archive. IEEE CAA J. Autom. Sinica (2019)
8. Dempster, A., Schmidt, D.F., Webb, G.I.: Minirocket: A very fast (almost) deterministic transform for time series classification. In: Proc. of KDD (2021)
9. Deng, G., Han, C., Dreossi, T., Lee, C., Matteson, D.S.: Ib-gan: A unified approach for multivariate time series classification under class imbalance. In: Proc. of SDM (2022)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of CVPR (2016)

11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. (1997)
12. Huang, H., Xu, C., Yoo, S., Yan, W., Wang, T., Xue, F.: Imbalanced time series classification for flight data analyzing with nonlinear granger causality learning. In: Proc. of CIKM (2020)
13. Jiang, Z., Chen, T., Chen, T., Wang, Z.: Improving contrastive learning on imbalanced data via open-world sampling. In: Proc. of NeurIPS (2021)
14. Kang, B., Li, Y., Xie, S., Yuan, Z., Feng, J.: Exploring balanced feature spaces for representation learning. In: Proc. of ICLR (2021)
15. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: Proc. of ICLR (2020)
16. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: Proc. of CVPR (2017)
17. Lee, D., Lee, S., Yu, H.: Learnable dynamic temporal pooling for time series classification. In: Proc. of AAAI (2021)
18. Li, G., Choi, B., Xu, J., Bhowmick, S.S., Chun, K., Wong, G.L.: Shapenet: A shapelet-neural network approach for multivariate time series classification. In: Proc. of AAAI (2021)
19. Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R.S., Indyk, P., Katabi, D.: Targeted supervised contrastive learning for long-tailed recognition. In: Proc. of CVPR (2022)
20. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. of ICCV (2017)
21. Reiss, A., Stricker, D.: Introducing a new benchmarked dataset for activity monitoring. In: 16th International Symposium on Wearable Computers, ISWC 2012, Newcastle, United Kingdom, June 18-22, 2012 (2012)
22. Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., Li, H.: Balanced meta-softmax for long-tailed visual recognition. In: Proc. of NeurIPS (2020)
23. Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. In: Proc. of CVPR (2021)
24. Wang, P., Han, K., Wei, X., Zhang, L., Wang, L.: Contrastive learning based hybrid networks for long-tailed image classification. In: Proc. of CVPR (2021)
25. Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., Xu, H.: Time series data augmentation for deep learning: A survey. In: Proc. of IJCAI (2021)
26. Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., Xu, B.: Ts2vec: Towards universal representation of time series. In: Proc. of AAAI (2022)
27. Zang, Y., Huang, C., Loy, C.C.: FASA: feature augmentation and sampling adaptation for long-tailed instance segmentation. In: Proc. of ICCV (2021)
28. Zha, D., Lai, K.H., Zhou, K., Hu, X.: Towards similarity-aware time-series classification. In: Proc. of SDM (2022)
29. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. CoRR (2021)
30. Zhao, P., Luo, C., Qiao, B., Wang, L., Rajmohan, S., Lin, Q., Zhang, D.: T-smote: Temporal-oriented synthetic minority oversampling technique for imbalanced time series classification. In: Proc. of IJCAI (2022)
31. Zhou, B., Cui, Q., Wei, X., Chen, Z.: BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proc. of CVPR (2020)
32. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proc. of AAAI (2021)