Time Series Class-Incremental Learning via Confidence-guided Mask Distillation and Prototype-guided Contrastive Learning

Yu Liu[1], Haoqin Yang[1], Jinping Sui[2*], Hui Wang[1], Haipeng Li[1], Weimin Wang[1], Qi Jia[1*]

[1]Dalian University of Technology, Dalian, China
[2]Dalian Naval Academy, Dalian, China
{liuyu8824,jiaqi,wangweimin}@dlut.edu.cn, {yanghaoqin,wanghuiharu}@mail.dlut.edu.cn,
suijinping13@alumni.nudt.edu.cn, 1241566148@mail.dlut.edu.cn

Abstract

Class-incremental learning (CIL) has recently gained great attention in the field of time series classification. Existing CIL methods based on knowledge distillation exhibit impressive ability to retain prior knowledge and overcome catastrophic forgetting, however, their effectiveness faces major challenges posed by time series data. Since temporal data is more susceptible to sensor errors and electronic noise, the distillation process may be significantly affected by noisy knowledge transfer. To address this issue, we propose a novel confidence-guided mask distillation (CMD) framework, to prevent the noisy inheritance during distillation. The core of CMD lies in a dynamic masking mechanism guided by prediction confidence, capable of allocating higher weights to high-confidence time series and substantially suppressing the influence of low-confidence ones. Additionally, different from prior work simply passing a set of feature prototypes to the classifier, we develop prototype-guided contrastive learning (PCL) to alleviate the classifier bias on new classes, through extra contrastive constraints to push away the feature distributions of old feature prototypes from those of new classes features. Extensive experiments on three time-series datasets demonstrate that, our method significantly outperforms other replay-free CIL approaches in raising average accuracy, as well as decreasing forgetting rate.

Code — https://github.com/YangHaoqin/CMD-PCL

Introduction

Time-series data, as its records the evolution of variables over time, inherently exhibits strong temporal dependencies, distributional drifts, and unbounded streaming growth (Kaushik et al. 2020; Ji et al. 2022; Zaini et al. 2022; Wu et al. 2024). However, most mainstream deep learning methods assume static data acquisition and one-time training fashion, making them ill-suited to adapt to dynamic characteristics of multivariate time series. To enable machine learners capable of acquiring new knowledge continuously as what humans do, class-incremental learning (CIL) has gained significant attention due to its ability to learn newly arriving data across incremental sessions (Zhou et al. 2024).

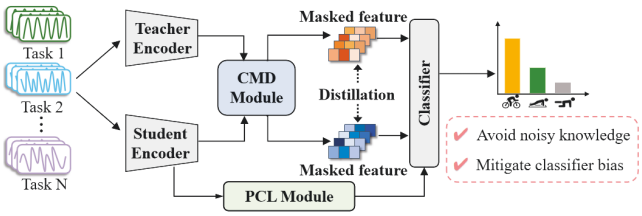


Figure 1: Conceptual scheme of our proposed framework for time series class-incremental learning (TSCIL). Different from prior work, our method constructs a confidence-guided mask distillation (CMD), to avoid noisy knowledge transfer from the teacher to the student. Besides, prototype-guided contrastive learning (PCL) is devised to mitigate the classifier bias on new classes.

To mitigate the so-called catastrophic forgetting encountered during incremental learning (French 1999), an effective remedy is to replay stored samples of old classes to preserve prior knowledge (Rolnick et al. 2019). However, this replay mechanism often raises data privacy concerns and is difficult to apply in resource-constrained scenarios due to its substantial storage overhead (Zhou et al. 2024).

Instead, extensive efforts primarily rely on two complementary strategies to accomplish replay-free class-incremental learning. The first is motivated by knowledge distillation (Lin et al. 2024), in which an old model (teacher) transfers its knowledge to a new model (student), to remember previous knowledge with respect to old classes. The second relies on feature prototypes (Zhu et al. 2021; Zhang et al. 2024a), which learns and memorizes one prototype for each old class (*i.e.* , class mean) to prevent forgetting. For instance, DT^{2}W (Qiao et al. 2023) integrates knowledge distillation and feature prototypes jointly. It uniformly distills the teacher's feature outputs at every time step, and aligns time-series shapes using dynamic time warping (DTW) to facilitate knowledge distillation. Meanwhile, prototype augmentation is further used to make the model retain the memory of old classes while learning new ones.

Nonetheless, the characteristics of time series data challenge these CIL methods significantly. Compared to image data, temporal data is more susceptible to sensor errors and electronic noise. However, the native knowledge distillation

*Corresponding authors.

often lacks reliability assessment and noise-filtering mechanisms. As a result, the student may be negatively affected by the uncertain knowledge distilled from the teacher. In addition, since the number of feature prototypes for old classes is significantly lower than that of new classes samples, such imbalanced data distribution makes the model easily overfit on new classes, thereby degrading the performance on old classes. Particularly when old and new classes are semantically similar, the prototypes tend to entangle closely with new classes in the latent feature space, leading to ambiguous and difficult-to-distinguish decision boundaries.

To address the noisy knowledge transfer from the teacher to the student, we propose a novel Confidence-guided Mask Distillation (CMD) framework for time series class-incremental learning (TSCIL). Concretely, to prevent the noisy inheritance during distillation, we devise a dynamic masking mechanism guided by prediction confidence. Unlike traditional uniform-weight distillation, we employ Monte Carlo Dropout to quantify the teacher's uncertainties, which are used to generate a dynamic mask, capable of allocating higher weights to high-confidence time series and substantially suppressing the influence of low-confidence ones. Through this masked knowledge distillation, our approach not only mitigates catastrophic forgetting, but also enhances the student's robustness to noisy information inherited from the teacher. Built upon CMD, we further develop Prototype-guided Contrastive Learning (PCL) to alleviate the classifier bias on new classes. Different from existing methods that simply feed a set of feature prototypes to the classifier, our PCL enforces extra contrastive constraints to push away the feature distributions of old feature prototypes from those of new classes features. It is beneficial for sharpening decision boundaries and reducing overfitting to new classes. At last, we integrate CMD and PCL and optimize them jointly in a unified model.

The main contributions of this paper are threefold:

- We address time series class-incremental learning via a novel confidence-guided masked distillation (CMD), which substantially improves the quality of temporal knowledge distillation and retention.
- We introduce prototype-guided contrastive learning (PCL), to effectively resolve the distribution overlap between old and new classes, as a result of alleviating the classifier bias toward new classes.
- We conduct comprehensive experiments on three popular multivariate time series datasets, where the results verify that our approach, consistently outperforming other replay-free competitors, raises the accuracy performance while decreasing the forgetting ratios remarkably.

Related Work

Class Incremental Learning

Class-incremental learning (CIL) allows a model to gradually learn new classes while retaining knowledge previously learned from old classes (Zhou et al. 2024). Its core challenge is caused by catastrophic forgetting, which means the performance on previously learned tasks drops significantly after new tasks are arriving. To prevent forgetting, CIL methods broadly fall into three categories: replay-based, regularization-based, and dynamic-architecture methods. Concretely, replay-based methods consolidate knowledge by replaying a few old samples (Rebuffi et al. 2017; Aljundi et al. 2019; Tiwari et al. 2022), while dynamic-architecture methods (Yan, Xie, and He 2021; Ermis et al. 2022; Douillard et al. 2022) allocate independent parameters for new tasks by expanding the network. Different from them, however, regularization-based methods protect old knowledge by imposing additional constraints while the student learns on new tasks. Such methods can be further divided into two mainstream directions: the first is parameter-constraint strategy that preserves knowledge by penalizing changes to parameters critical for old tasks (Aljundi, Kelchtermans, and Tuytelaars 2019; Lee et al. 2020; Yin, Yang, and Li 2021); the second is knowledge distillation strategy, which utilizes the old model as a teacher to compel the new model to maintain a consistent output with respect to new data (Li and Hoiem 2018; Douillard et al. 2020; Szatkowski et al. 2024). Moreover, emerging prompt-based incremental learning strategies (Zhang et al. 2024b), as a novel branch of regularization-based methods, leverage prompt engineering to adapt to new classes. However, such CIL methods often rely on two assumptions that are difficult to satisfy in real-world scenarios: the data distribution remains stable throughout the learning process, and clear task boundaries are available. However, for time series applications, the data typically evolves over time and arrives in a continuous streaming manner. Their statistical properties may drift due to environmental changes or noise perturbations. To address this issue, this work explores a new framework with the aim of relaxing these assumptions and achieving more robust incremental learning across diverse temporal domains.

Time Series Class Incremental Learning

Conventional time series classification methods (Wen et al. 2024; Wang et al. 2024; Chen et al. 2025) follow a static training paradigm, which inherently struggles to adapt to dynamic data streams where new classes emerge continuously. Focused on time series class-incremental learning (TSCIL), (Qiao et al. 2024) establish a unified evaluation benchmark, systematically compare regularization-based and replay-based methods. For instance, DT^2W (Qiao et al. 2023) utilizes soft dynamic time warping to align temporal shapes between old and new models. The methods such as CLOPS (Kiyasseh, Zhu, and Clifton 2021) and OCL-HAR (Schiemer et al. 2023), despite enhancing the performance through various manners, still rely on storing real data samples. To avoid reliance on real historical data, some studies in related fields have explored generative or hierarchical strategies. For instance, (Wang et al. 2019) tackles the forgetting problem in sound classification by training a generator for audio spectrograms. Likewise, HAD (Zuo et al. 2024) addresses class-incremental audio-visual recognition with a multi-level augmentation and distillation strategy. FCAC (Li et al. 2023) achieves few-shot incremental audio classification by adjusting class prototypes with self-attention. MAPIC (Sun et al. 2023) optimizes class prototypes via meta-learning, whereas it fails to address noise ro-

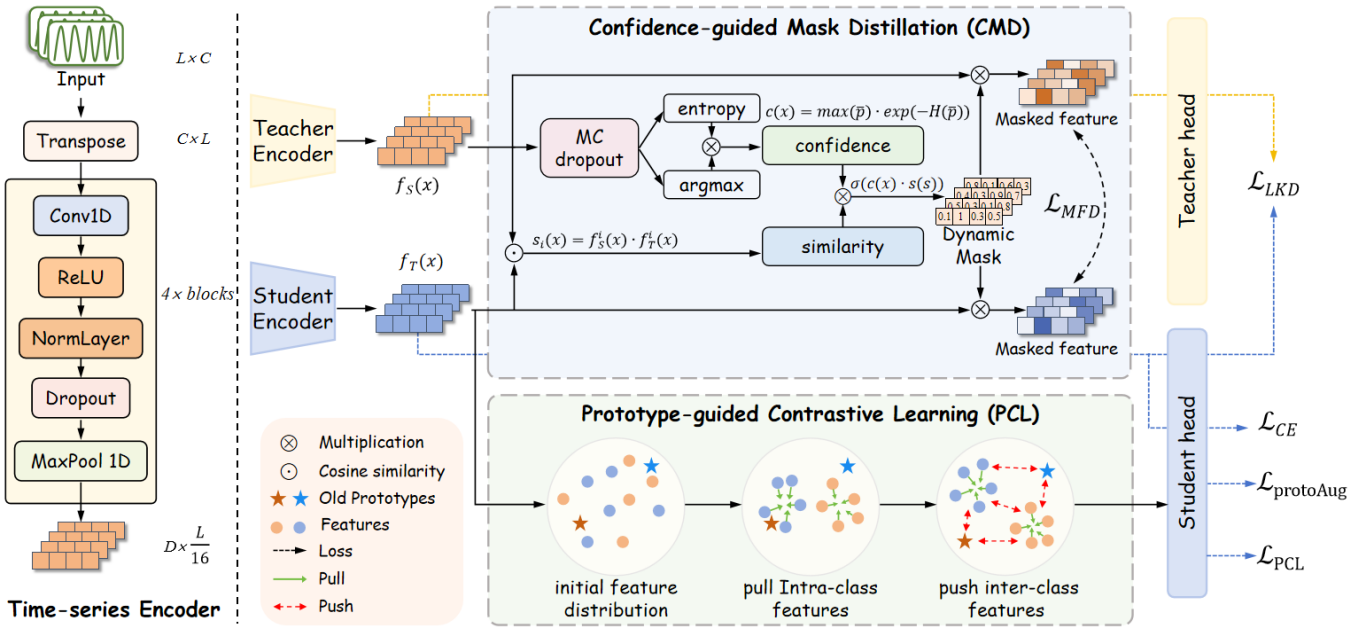


Figure 2: **Overview framework of our model for time series class-incremental learning.** After extracting temporal features from the teacher and student encoders, we pass them into the confidence-guided mask distillation (CMD) module, where a dynamic mask is generated to prevent the noisy inheritance during knowledge distillation. In addition, we develop prototype-guided contrastive learning (PCL), enforcing extra contrastive constraints to push away the feature distributions of old feature prototypes from those of new classes features, thereby alleviating the classifier bias on new classes. We optimize CMD and PCL jointly by minimizing several loss functions induced by the teacher and student heads.

bustness. We note that, despite the contributions of these methods, both noise robustness and classifier bias remain key bottlenecks when handling complex time series data. Inspired by this, our method suppresses noise via confidence-guided mask distillation, and reduces overfitting on new classes via prototype-guided contrastive learning.

Methodology

Problem Definition. The ultimate objective of TSCIL is to train a unified learner capable of maintaining accurate recognition on all time series classes observed up to the current task. Formally, we consider a sequence of B tasks $\mathcal{T} = \{T_1, T_2, \ldots, T_B\}$ where each task $T_b = (D_b, C_b)$ comprises a set of new classes $C_b = \{c_1, c_2, \ldots, c_{k_b}\}$ and its associated training dataset $D_b = \left\{\left(x_i^b, y_i^b\right)\right\}_{i=1}^{N_b}$, where N_b denotes the number of samples in task b, each $x_i^b \in \mathbb{R}^{l \times d}$ is a multivariate time series of and d variables over l time steps; $y_i^b \in C_b$ is the corresponding class label. TSCIL typically assumes that the class sets of different tasks are disjoint, which means $C_b \cap C_{b'} = \varnothing$, for any $b \neq b'$. Note that, we focus on a replay-free setting: when trained on task b, the model has access only to current data D_b, but cannot revisit data from prior tasks.

Confidence-guided Mask Distillation (CMD)

The proposed CMD enables the student model to prioritize learning high-confidence knowledge from the teacher model, thereby improving the effectiveness of knowledge distillation and retention. It comprises three key components including confidence estimation, dynamic mask generation, and masked feature distillation.

Confidence Estimation. To identify the regions where the teacher model possesses reliable knowledge, we quantify its predictive uncertainty using Monte Carlo (MC) Dropout (Gal and Ghahramani 2016). For a time series sample x, we perform $N = 15$ stochastic forward passes, producing a set of prediction probabilities $\{p_1, p_2, \ldots, p_N\}$. Next, we compute mean probability $\bar{p}$ and entropy $\mathcal{H}(\bar{p})$:

$$\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i, \mathcal{H}(\bar{p}) = -\sum_{c=1}^{C} \bar{p}_c \log(\bar{p}_c). \quad (1)$$

As a result, we estimate a new scalar, $c(x)$, which represents the teacher model's confidence for sample x:

$$c(x) = \max(\bar{p}) \cdot \exp(-\mathcal{H}(\bar{p})), \quad (2)$$

where $\max(\bar{p})$ ensures that the teacher outcome has a certain prediction, while $\exp(-\mathcal{H}(\bar{p}))$ imposes an exponential penalty on high-entropy predictions. We note that, $c(x)$ serves as a highly reliable estimation to guide the subsequent distillation process.

Dynamic Mask Generation. To prioritize reliable information and suppress noise, we construct a dynamic mask to enable confidence-guided selective weighting, thereby improving the robustness of knowledge transfer.

First, we capture temporal features, $f_S(x) \in \mathbb{R}^{L \times D}$ and $f_T(x) \in \mathbb{R}^{L \times D}$, from the student and teacher models, respectively. Then we permute them to $\mathbb{R}^{D \times L}$ and compute

the cosine similarity along the time dimension for each channel, which means $s_i(x) = \mathbf{f}_S^i(x) \cdot \mathbf{f}_T^i(x)$, where $\mathbf{s} \in \mathbb{R}^D$ is the resulting similarity vector, and $\mathbf{f}_S^i$ and $\mathbf{f}_T^i \in \mathbb{R}^L$ are the vectors for the i-th channel ($i = 1, \ldots, D$). At last, $\mathbf{s}(x)$ is multiplied by the confidence scalar $c(x)$, and their result is further passed through a Sigmoid function to obtain the final temporal mask:

$$\mathbf{M}(x) = \sigma(c(x) \cdot \mathbf{s}(x)). \tag{3}$$

The goal of dynamic mask $\mathbf{M}(x)$ is to guide the feature distillation process, by applying fine-grained attention to both the student and teacher feature maps. It ensures that the subsequent knowledge distillation process focuses more on high-confident information mutually shared by both models.

Masked Feature Distillation. Moreover, we apply the dynamic mask to the features of the student and teacher models, yielding two weighted feature maps: $f'_S(x) = \mathbf{M}(x) \odot f_S(x)$ and $f'_T(x) = \mathbf{M}(x) \odot f_T(x)$. Inspired by (Qiao et al. 2023), we employ a differentiable time-series alignment algorithm, namely soft dynamic time warping (Soft-DTW), to quantify masked feature distillation between $f'_S(x)$ and $f'_T(x)$:

$$\mathcal{L}_{\mathrm{MFD}} = \frac{1}{N^b} \sum_{i=1}^{N^b} \mathrm{SoftDTW}\left(f'_S(x_i), f'_T(x_i)\right). \tag{4}$$

where N^b denotes the number of time series samples in current task. Concretely, we detail the $\mathrm{SoftDTW}(\cdot)$ function as follows. First, we compute the pairwise squared Euclidean distances between the two weighted features to form a cost matrix $D_{i,j} = \|f'_S(x)_i - f'_T(x)_j\|_2^2$. Then we construct an accumulated cost matrix via dynamic programming:

$$R_{i,j} = D_{i,j} + \mathrm{softmin}\delta\left(Ri-1, j, R_{i-1,j-1}, R_{i,j-1}\right), \tag{5}$$

where $R_{i,0} = R_{0,j} = +\infty, R_{0,0} = 0$ controls boundary conditions. The operator $\mathrm{softmin}\delta$ replaces the non-differentiable $\min$ operator in classical DTW by:

$$\mathrm{softmin}\,\delta\left(a_1, a_2, a_3\right) = -\delta \log\left(\sum_{k=1}^{3} e^{-a_k/\delta}\right), \tag{6}$$

where the smoothing parameter δ, which is set to $\delta = 1.0$, controls the smoothness of the softmin approximation. As a result, the bottom-right element of the accumulated cost matrix R forms the Soft-DTW distance between the two weighted features, constituting the complete feature distillation loss $\mathcal{L}_{\mathrm{MFD}} = R_{N_S,N_T}$.

To complement feature knowledge distillation, we further adopt a logit-level knowledge distillation loss $\mathcal{L}_{\mathrm{LKD}}$:

$$\mathcal{L}_{\mathrm{LKD}} = \frac{1}{N^b} \sum_{i=1}^{N^b} \mathrm{KL}\left(P_T^{\mathrm{old}}(x_i) \parallel P_S^{\mathrm{old}}(x_i)\right), \tag{7}$$

where KL indicates the Kullback-Leibler divergence. $P_S^{\mathrm{old}}(\cdot)$ and $P_T^{\mathrm{old}}(\cdot)$ are the softened probabilities for the old classes derived from the student and teacher models, respectively. Ultimately, $\mathcal{L}_{\mathrm{FKD}}$ and $\mathcal{L}_{\mathrm{LKD}}$ are combined to form the complete distillation loss trained in our model:

$$\mathcal{L}_{\mathrm{KD}} = \lambda_1 \cdot \mathcal{L}_{\mathrm{MFD}} + \mathcal{L}_{\mathrm{LKD}}, \tag{8}$$

where λ_1 adjusts the balance between the two loss terms.

Prototype-guided Contrastive Learning (PCL)

Passively distilling knowledge alone is not sufficient to prevent forgetting, especially for complicated time series data. In addition to knowledge distillation, the use of feature prototypes can further mitigate the forgetting ratios. However, the imbalanced distribution between old classes prototypes and new classes samples leads to insufficient representativeness of old classes in the feature space, thereby making the model biased on new classes largely. Different from prior methods that simply pass feature prototypes into the classifier, we develop a prototype-guided contrastive learning mechanism, which can push apart the feature distributions of old prototypes and new samples, and sharpen the decision boundaries for alleviating forgetting on old classes. We elaborate the details in the following steps.

Prototype Construction. Feature prototype $P_c \in \mathbb{R}^d$, defined as the centroid embedding of class c, serves as a crucial global anchor. For old classes, we avoid using the yet-unrefined student model to generate prototypes, since this would introduce noise. Instead, before learning a new incremental task, we feed all training samples of the old tasks through the fully trained teacher model and take their mean feature as a static prototype: $P_c^{\mathrm{old}} = \frac{1}{|D_c|} \sum_{x_i \in D_c} f_T(x_i)$. For newly learned classes, we adopt an effective strategy: once the training of current incremental task completes, we switch the student model f_S to evaluation mode and extract features for all the training samples. We then compute each new class prototype as the arithmetic mean of its sample embeddings by $P_c^{\mathrm{new}} = \frac{1}{|D_c|} \sum_{x_i \in D_c} f_S(x_i)$. This procedure guarantees the stability and accuracy of feature prototypes.

Multi-level Structural Constraints. To overcome classifier bias in incremental learning, an ideal feature space should have three key properties: (a) intra-class compactness, to enhance robustness to intra-class variations; (b) global inter-class separation, to maintain discriminative boundaries between new and old classes; (c) local inter-class discriminability, to provide fine-grained distinction for easily confusable classes in current task. However, a single loss function cannot achieve these goals simultaneously. To this end, we design a hybrid contrastive loss $\mathcal{L}_{\mathrm{PCL}}$, which decomposes the overall objective into three independent sub-objectives and integrates them into a unified InfoNCE framework. First, regarding intra-class compactness, we encourage tight clustering by maximizing the similarity, $s_i^+ = \frac{1}{\tau} \mathbf{Z}_i^\top \mathbf{Z}_i^+$, between each batch feature $\mathbf{Z}_i$ (as the anchor) and a random same-class positive $\mathbf{Z}_i^+$. Second, for global inter-class separation, we push new features away from old clusters by contrasting $\mathbf{Z}_i$ with old prototypes $\{\pi_j\}_{j=1}^M$: $s_{i,j}^{\mathrm{proto}} = \frac{1}{\tau} \mathbf{Z}_i^\top \pi_j$. Third, to perform local inter-class separation, we define the negatives as features with different labels and contrast them with $\mathbf{Z}_i$. The original similarities are $S_{ik} = \frac{1}{\tau} \mathbf{Z}_i^\top \mathbf{Z}_k$, and we mask the same class via $\widetilde{S}_{ik} = S_{ik}$ if $y_k \neq y_i$, else $-\infty$. As a result, we combine the scores into a logit vector v_i:

$$v_i = \left[s_i^+; s_{i,1}^{\mathrm{proto}}, \ldots, s_{i,M}^{\mathrm{proto}}; \widetilde{S}_{i1}, \ldots, \widetilde{S}_{iN}\right] \in \mathbb{R}^{1+M+N}. \tag{9}$$

As a result, we compute the cross-entropy loss over v_i:

$$\mathcal{L}_{\mathrm{PCL}} = -\frac{1}{N^b}\sum_{i=1}^{N^b}\log\left(\mathrm{softmax}(v_i)\right). \tag{10}$$

In addition to the prototype-guided contrastive learning loss, we add a prototype augmentation loss to bolster the global structure in feature space. It helps to train the model classifying augmented old classes features correctly. Finally, the total prototype loss becomes:

$$\mathcal{L}_{\mathrm{Proto}} = \lambda_2 \cdot \mathcal{L}_{\mathrm{PCL}} + \mathcal{L}_{\mathrm{ProtoAug},} \tag{11}$$

where λ_2 balances the two terms.

Overall Objective Function

To jointly optimize knowledge distillation and feature prototype for time series class-incremental learning, we train the model incrementally with the following total loss cost:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{KD}} + \mathcal{L}_{\mathrm{Proto},} \tag{12}$$

where $\mathcal{L}_{\mathrm{CE}}$ indicates the basic cross-entropy loss, ensuring that the student model accurately learns new classes.

Experiments

Dataset Protocols. We conduct comprehensive experiments on three publicly available time-series datasets, UCI-HAR (Anguita et al. 2013), DSA (Altun, Barshan, and Tunçel 2010) and WISDM (Weiss 2019). The datasets are widely used in human activity and gesture recognition, and differ in acquisition devices, data dimensionality, sampling frequency, and activity complexity. Specifically, UCI-HAR comprises time-series data for 6 daily activities. Each sample consists of 9 channels over 128 time steps. DSA provides sensor recordings of 19 daily motion activities, containing 45 channels over 125 time steps. Notably, following the common practice in (Qiao et al. 2023), we select 12 classes from DSA for for class-incremental learning. WISDM contains 18 activities, with samples extracted by applying a non-overlapping sliding window with a window size of 200. Each sample includes 3 channels sampled at 20 Hz. Like DSA, 12 classes in WISDM are used for class-incremental learning as well. The data setups are summarized in Table 1.

Implementation Details. To ensure fairness, we adopt a widely used 1D-CNN as backbone. It comprises three stacked convolutional blocks, each consisting of a 1D convolutional layer, a ReLU activation, batch normalization, and max-pooling, followed by dropout. All models are trained from scratch using the Adam optimizer with an initial learning rate of 1.0 and a batch size of 64. We train the model for 100 epochs per incremental task, employing early stopping to monitor convergence. Experiments were conducted on one NVIDIA RTX 4090 GPU.

Evaluation Metrics. Following prior work (Chaudhry et al. 2018), we evaluate the models with two standard metrics: average accuracy A_T and average forgetting F_T. To establish thorough evaluation, we experiment with multiple runs and measure the average performance. For each run, we randomize the sequence of incremental tasks, to validate the generalization ability under varying class orders.

Dataset	Shape	Train Size	Test Size	# Tasks	# Classes per Task
UCI-HAR	9×128	7352	2947	3	2
DSA	45×125	6840	2280	6	2
WISDM	3×200	18184	6062	6	2

Table 1: Summary of dataset setups for TSCIL, including the shape of time series data, the number of training samples, the number of test samples, the number of incremental tasks, and the number of classes per task.

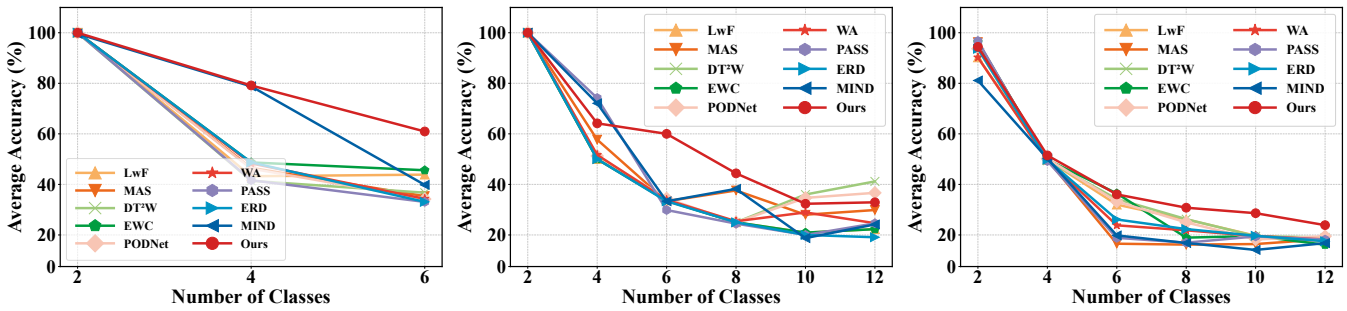
Method	UCI-HAR		DSA		WISDM	
	$A_T \uparrow$	$F_T \downarrow$	$A_T \uparrow$	$F_T \downarrow$	$A_T \uparrow$	$F_T \downarrow$
Joint training	93.9	N/A	99.5	N/A	85.3	N/A
Naive fine-tuning	32.9	97.8	17.6	98.8	15.5	95.6
EWC (PNAS'17)	50.3	71.2	24.8	85.9	13.2	75.5
LwF (TPAMI'18)	40.6	79.7	22.3	84.6	14.7	80.3
MAS (ECCV'18)	49.4	69.1	25.6	66.1	11.9	63.8
PODNet (ECCV'20)	44.4	74.4	24.9	82.8	15.9	65.5
WA (CVPR'20)	36.4	92.2	21.6	93.5	14.0	71.1
PASS (CVPR'21)	42.4	76.0	<u>28.4</u>	72.7	<u>18.1</u>	71.9
ERD (CVPR'22)	32.3	98.9	17.3	99.2	15.5	83.9
DT^{2}W (ICASSP'23)	<u>54.8</u>	<u>58.7</u>	24.5	87.2	16.4	<u>62.1</u>
MIND(AAAI'24)	35.7	75.7	25.6	<u>63.7</u>	11.6	65.4
Ours	**59.9**	**44.6**	**29.6**	**60.6**	**19.0**	**57.3**

Table 2: Comparison performance on three time series datasets. **Bold** numbers represent the best outcomes and the <u>underlined</u> ones as the second best.

Baseline Methods. To comprehensively validate the effectiveness of our method, we conduct comparative experiments against several state-of-the-art replay-free class-incremental learning approaches, including LwF (Li and Hoiem 2018), EWC (Kirkpatrick et al. 2017), MAS (Aljundi et al. 2018), WA (Zhao et al. 2020), PODNet (Douillard et al. 2020), PASS (Zhu et al. 2021), ERD (Feng, Wang, and Yuan 2022), MIND (Bonato et al. 2024). Although these compared methods are not designed for TSCIL initially, we carefully re-implement them with their publicly available codes and optimize their performance on three time series datasets. In addition, DT^{2}W (Qiao et al. 2023) is a representative competitor specifically designed for TSCIL. Apart from them, we implement two reference baselines: a joint training baseline that is trained with all classes in one time, representing the upper-bound performance; a naive fine-tuning baseline that trains new classes solely without using any forgetting-mitigation strategy.

Main Results

We summarize the results with respect to average accuracy (A_T) and average forgetting (F_T) in Table 2, where our method consistently outperforms the compared methods across three datasets and achieves new state-of-the-art performance. The following elaborates the compared results for three datasets: (1) On the UCI-HAR dataset, our method achieves 59.9% average accuracy and 44.6% average forgetting rate. Compared to the latest and state-of-the-art baseline DT^{2}W, our method improves accuracy by 5.1% and



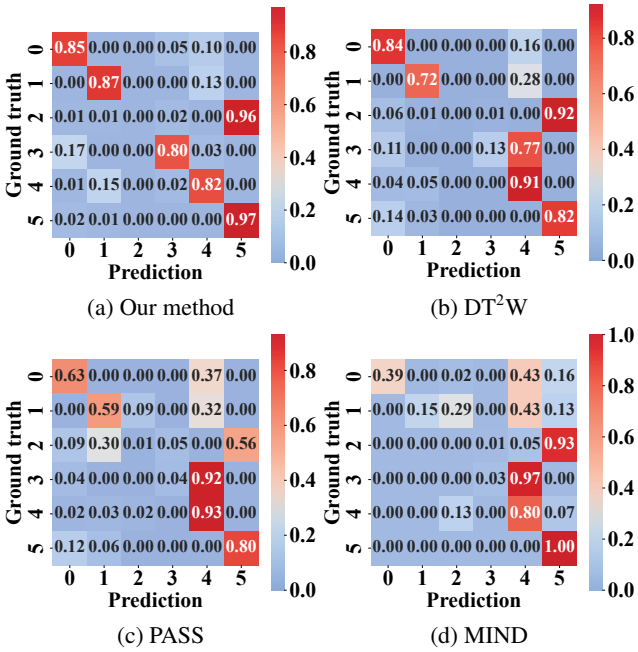
(a) UCI-HAR (b) DSA (c) WISDM

Figure 3: Accuracy curves during incrementally learning a sequence of tasks. It can be seen that, our method consistently surpasses other compared methods across incremental sessions for all three datasets.

reduces forgetting by 14.1%, showing a significant performance gain. (2) On the DSA dataset, our method attains 29.6% A_T and 60.6% F_T. Despite the higher difficulty in the DSA dataset, our method still improves DT^{2}W (24.5% accuracy, 87.2% forgetting) largely, boosting accuracy by 5.1% and reducing forgetting by 26.6%. Besides, our method surpasses the second-best methods PASS and MIND according to A_T and F_T, with accuracy raised by 1.2% and forgetting reduced by 3.1%, respectively. (3) On the WISDM dataset, our method obtains 19.0% average accuracy and 57.3% average forgetting rate. Although the large amount of samples in WISDM makes incremental learning more challenging, our method still holds the superiority over the baseline DT^{2}W (16.4% accuracy, 62.1% forgetting), increasing accuracy by 2.6% and decreasing forgetting by 4.8%. At the same time, our method surpasses the second-best method PASS (18.1% A_T), with a boost of 0.9%. We note that, these significant improvements are mainly attributed to our confidence-guided mask distillation and prototype-guided contrastive learning, effectively leveraging knowledge retention by filtering noise and alleviating the classifier bias on new classes. This work advances the research on solving the stability-plasticity dilemma in TSCIL.

Comparison with Accuracy Curves. To shed more light on the incremental learning paradigm, we further demonstrate the evolution of the accuracy along with new tasks arriving incrementally. It helps reflect the model's overall recognition ability for all previously seen tasks after learning each additional task. As shown in Fig. 3, it is evident that across all three datasets, despite the ongoing learning process, our average accuracy consistently remains at a higher level than other compared methods. Even after all tasks have been learned, the performance curves show only a slight decline. The results strongly support the effectiveness of our approach in overcoming catastrophic forgetting, as it effectively retains knowledge of prior tasks while flexibly acquiring new category knowledge, thereby achieving a promising balance between stability and plasticity.

Comparison with Confusion Matrix. To further examine the ability to distinguish knowledge between old and new classes, as well as the effectiveness in alleviating classifier bias, we compare the classification confusion matrix of our method with several baselines, after completing all incremental sessions. As shown in Fig. 4, the results signify that our method holds a significant advantage in correctly predicting old classes (*i.e.* , the upper-left region of each matrix), and at the same time, substantially mitigates the common classifier bias towards new classes in incremental learning, because of effectively separating old and new classes.



(a) Our method (b) DT^{2}W

(c) PASS (d) MIND

Figure 4: Confusion matrices of our method, DT^{2}W, PASS and MIND on the UCI-HAR dataset. Each cell represents the classification accuracy that is encoded according to the color bar. The darker color indicates better accuracy.

Ablation Study

We conduct extensive studies to validate the key components and critical hyperparameter settings used in our framework. The reported experiments below are conducted on the UCI-HAR dataset, with no loss of generality.

Baseline	CMD	PCL	$A_T \uparrow$	$F_T \downarrow$
✓			51.50	62.00
✓	✓		55.42	47.86
✓		✓	52.04	63.79
✓	✓	✓	**59.91**	**44.55**

λ_1 for $\mathcal{L}_{\text{MFD}}$	$A_T \uparrow$	$F_T \downarrow$
0.001	33.6	93.1
0.01	35.7	89.1
0.1	**59.8**	47.8
1.0	44.9	**46.8**

λ_2 for $\mathcal{L}_{\text{PCL}}$	$A_T \uparrow$	$F_T \downarrow$
0.01	58.3	56.2
0.1	**59.8**	**47.8**
0.5	54.9	60.6
1.0	48.5	69.9

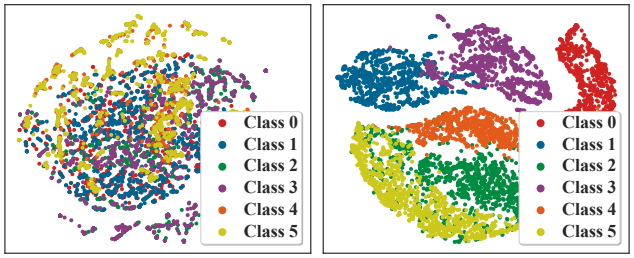
Table 3: Ablation study on the proposed components with the HAR dataset. **Left:** the component ablation of CMD and PCL. **Middle:** varying the trade-off parameter λ_1 for $\mathcal{L}_{\text{MFD}}$. **Right:** varying the trade-off parameter λ_2 for $\mathcal{L}_{\text{PCL}}$.

Component Analysis. To evaluate the individual contributions of each component in our method, we conduct ablation studies using DT^{2}W as the baseline. Given the ablation results reported in Table 3, we can see that, incorporating CMD yields the most significant improvements. Compared to the baseline, A_T increases by 3.92%, and F_T decreases by 14.14%, confirming the critical role of our confidence-guided mask distillation mechanism in stably preserving prior knowledge. In addition, as PCL effectively mitigates feature overlap between old and new classes and addresses the classification bias issue, thereby it boosts A_T by 0.54%, highlighting its key value in enhancing the model's discriminative capabilities. Ultimately, integrating both CMD and PCL achieves the largest performance gains. The full model realizes an 8.41% increase in the accuracy and a 17.45% reduction in the forgetting rate relative to the baseline. This comprehensive enhancement far surpasses the benefits of any single module, underscoring the powerful synergy among the components.

Hyperparameter Analysis. We conduct a series of comprehensive hyperparameter analyses to validate their impact on our performance. In particular, we mainly tune the two new hyperparameters, λ_1 and λ_2. The former is the loss weight for $\mathcal{L}_{\text{MFD}}$, and the later for $\mathcal{L}_{\text{PCL}}$. The results, as shown in Table 3, clearly reveal the decisive impact of both hyperparameters. Concretely, reducing λ_1 weakens knowledge distillation from the teacher model, leading to catastrophic forgetting of old tasks (with a forgetting rate exceeding 90%). Conversely, an excessively large λ_1 causes the model to over-rely on the teacher, limiting its ability to learn new classes and significantly harming the accuracy. Regarding λ_2, a value that is too low results in insufficient separation between old and new class features. In contrast, a value that is too high overly compresses the feature space and affects representation ability. Both scenarios ultimately hinder the acquisition of new knowledge. After a thorough tuning, when $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$, the model achieves optimal results in both accuracy and resistance to forgetting.

Qualitative Evaluation

To further validate the effectiveness of our prototype-guided contrastive learning on alleviating feature confusion, we apply *t*-SNE algorithm to visualize the feature distributions with the UCI-HAR dataset, as shown in Fig. 5. We can observe that, without $\mathcal{L}_{\text{PCL}}$, the feature space of time series data exhibits severe overlap between old and new classes, along with indistinct decision boundaries, which directly leads to more severe forgetting. In contrast, the model incorporating $\mathcal{L}_{\text{PCL}}$ successfully organizes the features of different classes into well-separated and compact clusters. This remarkable improvement in inter-class separability and intra-class compactness provides convincing evidence of $\mathcal{L}_{\text{PCL}}$ alleviating feature confusion and mitigating classifier bias towards new classes. This advantage is achieved by constructing a more discriminative representation space through the imposition of contrastive constraints. This qualitative evaluation corroborates our superior performance observed in the quantitative evaluations above.



(a) without $\mathcal{L}_{\text{PCL}}$ (b) with $\mathcal{L}_{\text{PCL}}$

Figure 5: Comparison of feature distributions between without and with $\mathcal{L}_{\text{PCL}}$ on the UCI-HAR dataset.

Conclusion

In this work, we have proposed a new framework that aims to better balance stability and plasticity in time series class-incremental learning (TSCIL). The framework is enhanced mainly through two components: confidence-guided mask distillation (CMD) and prototype-guided contrastive learning (PCL). CMD is introduced to reduce noisy knowledge transfer during distillation. It generates a dynamic confidence-aware mask that gives higher weights to reliable, high-confidence time series while suppressing the influence of low-confidence ones. We further develop PCL pushing away the feature distributions of old feature prototypes from those of new classes features, as a result of alleviating the classifier bias on new classes. Extensive results validate the superiority of our method over other competitive replay-free methods. Future work will investigate how to apply large time series models to TSCIL.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants No.62472066, No.42450226, No.62272083, No.62401602 and No.62306059.

References

Aljundi, R.; Babiloni, F.; Elhoseiny, M.; Rohrbach, M.; and Tuytelaars, T. 2018. Memory aware synapses: Learning what (not) to forget. In *Proc. of the European Conference on Computer Vision (ECCV)*, 139–154.

Aljundi, R.; Kelchtermans, K.; and Tuytelaars, T. 2019. Task-free continual learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 11254–11263.

Aljundi, R.; Lin, M.; Goujaud, B.; and Bengio, Y. 2019. Gradient Based Sample Selection for Online Continual Learning. In *Advances in Neural Information Processing Systems*, 1058. Red Hook, NY, USA: Curran Associates, Inc.

Altun, K.; Barshan, B.; and Tunçel, O. 2010. Comparative study on classifying human activities with miniature inertial and magnetic sensors. 43(10): 3605–3620.

Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J. L.; et al. 2013. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, 3–4.

Bonato, J.; Pelosin, F.; Sabetta, L.; and Nicolosi, A. 2024. MIND: Multi-Task Incremental Network Distillation. In *Proc. of AAAI Conference on Artificial Intelligence*, volume 38, 11105–11113.

Chaudhry, A.; Dokania, P. K.; Ajanthan, T.; and Torr, P. H. S. 2018. Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence. In *Proc. of the European Conference on Computer Vision (ECCV)*, volume 11220, 556–572.

Chen, J.; Cao, T.; Xu, J.; Li, J.; Chen, Z.; Xiao, T.; and Yang, Y. 2025. Con4m: Context-Aware Consistency Learning Framework for Segmented Time Series Classification. In *Advances in Neural Information Processing Systems*, 3490–3519.

Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; and Valle, E. 2020. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Proc. of the European Conference on Computer Vision (ECCV)*, 86–102. Berlin, Heidelberg: Springer-Verlag.

Douillard, A.; Ramé, A.; Couairon, G.; and Cord, M. 2022. Dytox: Transformers for continual learning with dynamic token expansion. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 9285–9295.

Ermis, B.; Zappella, G.; Wistuba, M.; Rawal, A.; and Archambeau, C. 2022. Memory efficient continual learning with transformers. In *Advances in Neural Information Processing Systems*, volume 35, 10629–10642. Red Hook, NY, USA: Curran Associates Inc.

Feng, T.; Wang, M.; and Yuan, H. 2022. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 9427–9436.

French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4): 128–135.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proc. of the International Conference on Machine Learning*, 1050–1059.

Ji, J.; Wang, J.; Jiang, Z.; Jiang, J.; and Zhang, H. 2022. STDEN: Towards physics-guided neural networks for traffic flow prediction. In *Proc. of AAAI Conference on Artificial Intelligence*, volume 36, 4048–4056.

Kaushik, S.; Choudhury, A.; Sheron, P. K.; Dasgupta, N.; Natarajan, S.; Pickett, L. A.; and Dutt, V. 2020. AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in big data*, 3: 4.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Kiyasseh, D.; Zhu, T.; and Clifton, D. A. 2021. A clinical deep learning framework for continually learning from cardiac signals across diseases, time, modalities, and institutions. *Nature Communications*, 12: 4221.

Lee, J.; Hong, H. G.; Joo, D.; and Kim, J. 2020. Continual learning with extended kronecker-factored approximate curvature. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 9001–9010.

Li, Y.; Cao, W.; Xie, W.; Li, J.; and Benetos, E. 2023. Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes. 26: 1346–1360.

Li, Z.; and Hoiem, D. 2018. Learning without forgetting. 40(12): 2935–2947.

Lin, J.; Wu, Z.; Lin, W.; Huang, J.; and Luo, R. 2024. M2sd: Multiple mixing self-distillation for few-shot class-incremental learning. In *Proc. of AAAI Conference on Artificial Intelligence*, volume 38, 3422–3431.

Qiao, Z.; Hu, M.; Jiang, X.; Suganthan, P. N.; and Savitha, R. 2023. Class-incremental learning on multivariate time series via shape-aligned temporal distillation. In *Proc. of IEEE Conference on Acoustics, Speech and Signal Processing*, 1–5.

Qiao, Z.; Pham, Q.; Cao, Z.; Le, H. H.; Suganthan, P. N.; Jiang, X.; and Ramasamy, S. 2024. Class-incremental learning for time series: Benchmark and evaluation. In *Proc. of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 5613–5624.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. iCaRL: Incremental Classifier and Representation Learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 5533–5542.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T.; and Wayne, G. 2019. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32. Red Hook, NY, USA: Curran Associates Inc.

Schiemer, M.; Fang, L.; Dobson, S.; and Ye, J. 2023. Online continual learning for human activity recognition. *Pervasive and Mobile Computing*, 93: 101817.

Sun, L.; Zhang, M.; Wang, B.; and Tiwari, P. 2023. Few-shot class-incremental learning for medical time series classification. *IEEE Journal of Biomedical and Health Informatics*, 28(4): 1872–1882.

Szatkowski, F.; Pyla, M.; Przewięźlikowski, M.; Cygert, S.; Twardowski, B.; and Trzciński, T. 2024. Adapt your teacher: Improving knowledge distillation for exemplar-free continual learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 1977–1987.

Tiwari, R.; Killamsetty, K.; Iyer, R.; and Shenoy, P. 2022. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 99–108.

Wang, Y.; Huang, N.; Li, T.; Yan, Y.; and Zhang, X. 2024. Medformer: A multi-granularity patching transformer for medical time-series classification. In *Advances in Neural Information Processing Systems*, volume 37, 36314–36341.

Wang, Z.; Subakan, C.; Tzinis, E.; Smaragdis, P.; and Charlin, L. 2019. Continual learning of new sound classes using generative replay. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 308–312.

Weiss, G. M. 2019. Wisdm smartphone and smartwatch activity and biometrics dataset. *UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set*, 7(133190-133202): 5.

Wen, Y.; Ma, T.; Weng, L.; Nguyen, L.; and Julius, A. A. 2024. Abstracted shapes as tokens-a generalizable and interpretable model for time-series classification. In *Advances in Neural Information Processing Systems*, volume 37, 92246–92272.

Wu, H.; Liang, Y.; Xiong, W.; Zhou, Z.; Huang, W.; Wang, S.; and Wang, K. 2024. Earthfarsser: Versatile spatio-temporal dynamical systems modeling in one model. In *Proc. of AAAI Conference on Artificial Intelligence*, volume 38, 15906–15914.

Yan, S.; Xie, J.; and He, X. 2021. Der: Dynamically expandable representation for class incremental learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 3014–3023.

Yin, H.; Yang, P.; and Li, P. 2021. Mitigating forgetting in online continual learning with neuron calibration. In *Advances in Neural Information Processing Systems*, volume 34, 10260–10272.

Zaini, N.; Ean, L. W.; Ahmed, A. N.; and Malek, M. A. 2022. A systematic literature review of deep learning neural network for time series air quality forecasting. *Environmental science and pollution research international*, 29(4): 4958—4990.

Zhang, Y.; Jia, Q.; Fan, X.; Liu, Y.; and He, R. 2024a. CSC-Net: Class-Specified Cascaded Network for Compositional Zero-Shot Learning. In *Proc. of IEEE Conference on Acoustics, Speech and Signal Processing*, 3705–3709.

Zhang, Y.; Qiu, B.; Jia, Q.; Liu, Y.; and He, R. 2024b. Not Just Object, But State: Compositional Incremental Learning without Forgetting. In *Advances in Neural Information Processing Systems*.

Zhao, B.; Xiao, X.; Gan, G.; Zhang, B.; and Xia, S.-T. 2020. Maintaining discrimination and fairness in class incremental learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 13208–13217.

Zhou, D.-W.; Wang, Q.-W.; Qi, Z.-H.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2024. Class-Incremental Learning: A Survey. 46(12): 9851–9873.

Zhu, F.; Zhang, X.-Y.; Wang, C.; Yin, F.; and Liu, C.-L. 2021. Prototype Augmentation and Self-Supervision for Incremental Learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 5871–5880.

Zuo, Y.; Yao, H.; Zhuang, L.; and Xu, C. 2024. Hierarchical Augmentation and Distillation for Class Incremental Audio-Visual Video Recognition. 46(11): 7348–7362.