Isolation and Impartial Aggregation: A Paradigm of Incremental Learning without Interference

Yabin Wang[1, 3*], **Zhiheng Ma**[2*], **Zhiwu Huang**[3,4], **Yaowei Wang**[5], **Zhou Su**[1], **Xiaopeng Hong**[6, 5†]

[1] School of Cyber Science and Engineering, Xi'an Jiaotong University
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3] Singapore Management University
[4] University of Southampton
[5] Peng Cheng Laboratory
[6] Harbin Institute of Technology
iamwangyabin@stu.xjtu.edu.cn, zh.ma@siat.ac.cn, zhiwu.huang@soton.ac.uk, wangyw@pcl.ac.cn, zhousu@ieee.org,
hongxiaopeng@ieee.org

Abstract

This paper focuses on the prevalent stage interference and stage performance imbalance of incremental learning. To avoid obvious stage learning bottlenecks, we propose a new incremental learning framework, which leverages a series of stage-isolated classifiers to perform the learning task at each stage, without interference from others. To be concrete, to aggregate multiple stage classifiers as a uniform one impartially, we first introduce a temperature-controlled energy metric for indicating the confidence score levels of the stage classifiers. We then propose an anchor-based energy self-normalization strategy to ensure the stage classifiers work at the same energy level. Finally, we design a voting-based inference augmentation strategy for robust inference. The proposed method is rehearsal-free and can work for almost all incremental learning scenarios. We evaluate the proposed method on four large datasets. Extensive results demonstrate the superiority of the proposed method in setting up new state-of-the-art overall performance. *Code is available at* https://github.com/iamwangyabin/ESN.

Introduction

Incremental learning (*a.k.a*, continual learning or lifelong learning) is a paradigm that continually evolves machine learning models on a data stream. It is a longstanding research topic and might offer a path toward more human-like AI. The stability-plasticity dilemma is central to incremental learning (Mai et al. 2022; De Lange et al. 2021), which requires the models to be plastic to acquire new knowledge and stable to consolidate existing knowledge continuously.

Most previous works struggle to keep a fragile balance between stability and plasticity and also achieve pretty good results in terms of average accuracy, which, however, result in tremendous performance gaps of different learning stages (a.k.a. sessions or tasks). This is a well-known phenomenon named class imbalance (Mai et al. 2022; De Lange et al. 2021), which eventually leads to *stage performance imbalance*. That is, the newer the stage, the higher the performance, and vice versa. The source of this problem is twofold: firstly, the imbalance in the number of samples between the new incoming data and the historical data; secondly and more importantly, using a uniform model to portray a heterogeneous data stream may suffer from stage interference (Fig. 1 (a)) and lead to *a zero-sum game* (Riemer et al. 2018; Knoblauch, Husain, and Diethe 2020; Wang, Huang, and Hong 2022) in the performance of different stages. This results in breakdowns in recognizing certain classes, which creates a bottleneck in the final performance and limits the application of the model in real-world scenarios. A few methods use a rehearsal buffer to alleviate such imbalance problems (Hou et al. 2019). However, saving previous training data is memory expensive and has a privacy issue.

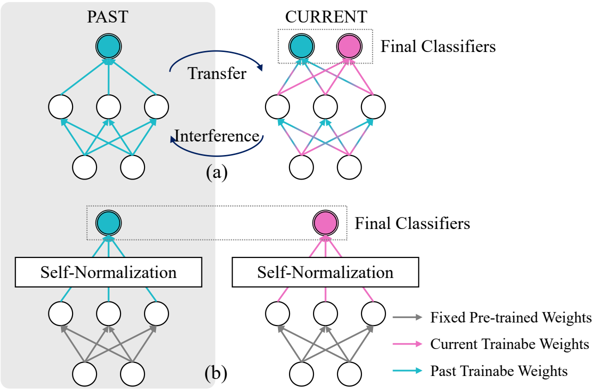

Figure 1: Comparison of our method against traditional methods. (a) Existing methods usually use a unified model for incremental learning, which may cause interference among the stages, as indicated by the lines of mixed colors. (b) In contrast, the proposed ESN uses a stage-isolation scheme for learning stage classifiers upon a fixed pre-trained backbone, resulting in much less forgetting and interference.

[*] These authors contributed equally.

[†] Xiaopeng Hong is the corresponding author.

To address these issues, in this paper, we study how to create *win-win* solutions for all-stage learning. We challenge the traditional unified paradigm and suggest a stage-isolation scheme for learning stage classifiers (Fig. 1 (b)). *Stage isolation* is targeted at learning multiple high confidence and low bias stage-specific classifiers for every stage in isolation, so that the classifier in each stage can be shielded from the interference of other stages, to satisfy the performance requirements of each stage adequately.

Nevertheless, the main difficulty of this paradigm lies in how to aggregate in an *impartial* way multiple isolated learners trained on different stages, as the learners trained on different stages of various incoming data stream separately may have diverse class-wise confidence distributions. For example, as shown in Fig. 3 (b), the output scores of the classifiers of two stages can be clearly different. A straightforward aggregation like finding the class with the highest confidence still has a tendency to *stage imbalance*.

We contend that the key to solving this problem lies in the regularization of the stage classifier outputs. Specifically, there are three criterions to meet. **Criterion 1**: each stage classifier should have a higher output confidence score for the data within the stage it belongs to (*i.e.*, in-stage data) than others (*i.e.*, out-stage data); **Criterion 2**: the confidence scores for in-stage data should be consistent across all stages; **Criterion 3**: the right stage classifier for the in-stage data shall have the highest confidence score among all classifiers. Unfortunately, they are challenging to be satisfied in the incremental learning scenarios. The reasons are two-fold. First, optimizing the stage learners only using the current data (the only one accessible) will result in a serious bias. Second, it is impossible to make full regularization as the classifiers to be learned in the future can not be considered at the current stage.

Dealing with such a *backward-compatible regularization* dilemma, inspired by the Helmholtz free energy theory (LeCun et al. 2006), we introduce a temperature-controlled energy metric to reflect the confidence score levels of stage classifiers. On this basis, we provide a *rehearsal-free* incremental learning paradigm, which regularizes the stage classifiers for aggregating them impartially as a uniform classifier. Specifically, we first use a frozen pre-trained ViT (Dosovitskiy et al. 2021) backbone as a strong prior before the stage-specific classifiers to assure higher confidence scores for the in-stage data than the out-stage data as far as possible, which has been proved in (Fort, Ren, and Lakshminarayanan 2021) (for **Criterion 1**). Second, we design an anchor-based energy self-normalization loss, which restricts the energy metrics of stage classifiers tightly around the energy anchor, to ensure all stage classifiers lay in the same energy level when facing in-stage data of their own (for **Criterion 2**). Furthermore, though the 'so far' best control parameter for the current stage can be found by a design method, it works only in a backward-compatible manner. To avoid overfitting to any specific stage, we maintain the 'so far' best parameters for all stages met and use a voting scheme to produce reliable inference outputs, by which **Criterion 3** can be better approached by stages[1].

To summarize, we propose a brand-new rehearsal-free general incremental learning paradigm to tackle the stage interference and stage performance imbalance problems, called **E**nergy **S**elf-**N**ormalization (**ESN**), which can handle almost all continual learning scenarios, including class-incremental learning (CIL) (De Lange et al. 2021), domain-incremental learning (DIL) (Wang, Huang, and Hong 2022), and cross-domain class incremental learning (Xie, Yan, and He 2022). The contributions can be further detailed as follows:

- We propose the anchor-based energy self-normalization (ESN) strategy so that stage-classifiers can produce high and consistent confidence scores for in-stage data.
- We design a control parameter (temperature) finding method to obtain stage-cumulative best parameters for progressively ensuring 'right' classifiers with the highest scores. On this basis, we propose a voting-based inference augmentation method for robust inference.
- The proposed ESN sets up new state-of-the-art performance, as shown by extensive experiments on four large-scale benchmark datasets. A challenging benchmark for cross-domain class incremental learning is built as well.

Related Works

Incremental Learning. There are three main types of incremental learning methods (De Lange et al. 2021).

Rehearsal-based methods tackle catastrophic forgetting either by keeping a small set of old training examples in memory (Tao et al. 2020a; Dong et al. 2021; Liu et al. 2022; Yang et al. 2022a,b) or using synthesized data produced by generative models (Shin et al. 2017). By using the rehearsal buffer for knowledge distillation and regularization, rehearsal-based methods have achieved state-of-the-art results on various benchmarks (Douillard et al. 2022; Joseph et al. 2022; Zhang et al. 2022). However, the performance of rehearsal-based methods generally deteriorates with a smaller buffer size (Mai et al. 2022). Moreover, it is desirable that the exemplars of old tasks are not stored for data security and privacy (Wang, Huang, and Hong 2022).

Regularization-based methods design knowledge distillation strategies (Li and Hoiem 2017; Tao et al. 2020b) or parameter regularization terms (Kirkpatrick et al. 2017) to mitigate catastrophic forgetting.

Network-based methods modify networks' architecture during incremental learning to mitigate catastrophic forgetting. Some works expand network parameters to learn new tasks and get state-of-the-art performances (Yan, Xie, and He 2021; Wang et al. 2022c; Xu and Zhang 2020; Douillard et al. 2022). Also some methods use the parameter isolation strategy to keep each task independent (Serra et al. 2018; Li et al. 2019). Recently, L2P (Wang et al. 2022c) and DualPrompt(Wang et al. 2022b) use prompt tuning and pre-trained models for incremental learning tasks. Prompt tuning offers an efficient way of model tuning for incremental learning. However, L2P and DualPrompt still use a

[1] **Criteria 1** and **2** together form a rough guarantee of **Criterion 3**, as shown in the 'Proposed Method' Section.

unified-structure model, and they require a fixed query function for task identification during inference, which is time-consuming and less effective in complicated situations.

Energy-based Models. Energy-based Models (EBMs) (LeCun et al. 2006) capture dependencies of variables by associating a scalar energy to each configuration of the variables. EBMs have been used for generative modeling (Du and Mordatch 2019; Xie et al. 2016), out-of-distribution detection (Liu et al. 2020), and open-set classification (Al Rahhal et al. 2022). Despite being successful across various tasks, EBMs have limited applications in incremental learning. ELI (Joseph et al. 2022) proposes to learn an energy manifold to counter the representational shift that happens during incremental learning. It uses EBMs to portray changes in the model and then tries to compensate the updated model to the original one, which is still a tug-of-war. What's more, it assumes the energy between different stages is distinguishable, which is a *too* strong assumption in application scenarios. EA (Zhao et al. 2022) also uses the energy-based model to add the calculated shift scalars onto the output logits to mitigate class imbalance. The calculation of compensation scalars is based on the samples of all classes, which suggests that it relies on a rehearsal buffer. Both works still struggle to alleviate the imbalance problem in a uniform model. Moreover, they are both rehearsal-based and can only handle CIL problems, which are far from general and robust solutions for incremental learning.

Proposed Method

Problem Definition. Incremental learning requests training a given machine learning model along a data stream, while the model can only access part of the training data at a time. Let $\zeta = \{1, 2, 3, ..., S\}$ denote the Stage-ID set, where S is the current maximum stage number. The incoming data of the s-th stage is denoted as $\mathcal{D}^s = \{x_i, y_i\}_{i=1}^{N^s}$, where N^s is the total sample number of this stage. $(x, y) \sim p_{data}^s$ represents the data distribution of the s-th stage. For class incremental learning, different stages have different categories to learn, and there is no category overlap, i.e., $\mathcal{Y}^i \cap \mathcal{Y}^j = \emptyset$, where $\mathcal{Y}^s$ is the label set of the s-th stage. For domain incremental learning, the categories maintains the same for all stages, $\mathcal{Y}^i = \mathcal{Y}^j$, but data distribution of each stage is different or even highly heterogeneous.

Our proposed ESN can handle these two challenging scenarios at the same time and even more challenging cross-domain class incremental learning, in which different stages have different categories from different domains.

Overall Framework. Previous incremental learning methods need to find a fragile balance between stability and plasticity. Using a unified model to portray a heterogeneous data stream may result in a zero-sum game and be seriously biased toward newer classes (Mai et al. 2022).

In this paper, we proposed a rehearsal-free general incremental learning paradigm to tackle the imbalance and the zero-sum game problems. Specifically, we train multiple isolated stage-specific classifiers upon a frozen pre-trained backbone for each stage. In the inference phase, we first select the most confident classifier (Eq. 1), and then use it to predict the final result (Eq. 2).

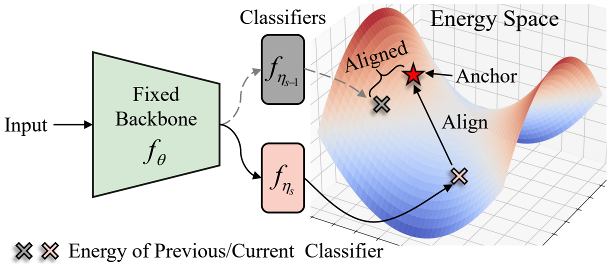

Figure 2: Overview of the proposed anchor-based energy self-normalization for stage classifiers. The classifiers of the current and the previous stages, f_{η_s} and $f_{\eta_{s-1}}$, are aligned sequentially by restricting their energies around the anchor.

$$s^* = \underset{s \in \zeta}{\operatorname{argmax}}\, H^s(x), \tag{1}$$

$$y^* = \underset{y \in \mathcal{Y}^{s^*}}{\operatorname{argmax}}\, P^{s^*}(y|x), \tag{2}$$

where $P^s(y|x)$ is the classifier of the s-th stage, and its confidence score function is denoted as $H^s(x)$. As shown, the comparability between different stages' confidence scores is the guarantee of impartial aggregation.

As shown in Fig. 1, given a pre-trained backbone f_θ, we initialize a specific classifier f_{η_s} in each stage. During training at stage s, we freeze the pre-trained backbone f_θ and only update the parameters of the classifier η_s. θ and η are parameters of backbone and classifier respectively. For simplicity, we use ViT-B/16 pre-trained on ImageNet as freeze backbone, and use the class-attention block (CAB) (Touvron et al. 2021) with a linear projection for the classifier. Our proposed strategy is also suitable for other parameter isolation methods like (Jia et al. 2022), which we will show later in the experiments. The stage isolated classifier can shield the interference of each other.

In the following sections, we first detail the training method based on the self-normalization strategy, which ensures the impartial aggregation of all stages' classifiers. Then we introduce the stage-cumulative control parameter optimization method with voting-based inference augmentation to further promote the performance.

Stage Classifier Self-Normalization. The most commonly used training criterion in training deep neural networks is the softmax cross-entropy loss. However, previous works (Tang et al. 2021; Liu et al. 2020) show that directly training with this loss results in overconfidence issues, where the maximum softmax activation value always approaches one in despite of whether the data is from training data distribution or not. Previous works have shown that other criteria such as the Helmholtz free energy (Liu et al. 2020) or the maximum logit value (Hendrycks et al. 2019) are better confidence scores than the maximum softmax value. However, **none of the above works discuss how to align confidence scores between different classifiers learned from data steam**.

Next, we first briefly review the relationship between the softmax cross-entropy loss and the energy-based model (Grathwohl et al. 2019; Liu et al. 2020; LeCun et al. 2006), then propose the anchor-based energy self-normalization objective function, which makes energy for in-stage data consistent across stages.

Let's define the energy function for a given input-label pair (x, y) as follows:

$$E^s(x, y) = -h^s(x)[y], \tag{3}$$

where $h^s(x) = f_{\eta_s}(f_\theta(x))$ is the logits of the s-th classifier, and $h^s(x)[y]$ is the logit value of $y \in \mathcal{Y}^s$, then softmax activation can be considered as a special case of discrete Gibbs distribution when the temperature parameter T equals to 1:

$$P_T^s(y|x) = \frac{\exp(-E^s(x, y)/T)}{\exp(-\mathcal{F}_T^s(x)/T)}, \tag{4}$$

where $\mathcal{F}_T^s(x)$ is the Helmholtz free energy, which can be expressed as the negative log partition function:

$$\mathcal{F}_T^s(x) = -T \log \sum_{y \in \mathcal{Y}^s} \exp\left(-E^s(x, y)/T\right). \tag{5}$$

Thus, the softmax cross-entropy loss is Eq. 6.

$$\begin{aligned} \mathcal{L}_{ce}^s &= \mathbb{E}_{(x,y)\sim p_{data}^s}(-\log P_T^s(y|x)) \\ &= \frac{1}{T}\mathbb{E}_{(x,y)\sim p_{data}^s}(E^s(y, x) - \mathcal{F}_T^s(x)). \end{aligned} \tag{6}$$

As can be seen, the softmax cross-entropy loss will decrease the energy between the input data and the ground-truth label while increasing the overall Helmholtz free energy. However, when $E^s(y, x)$ and $\mathcal{F}_T^s(x)$ are added with the same scalar, the loss value remains unchanged, which makes it meaningless to directly compare the free energy between different classifiers trained independently using softmax cross-entropy loss. To fix this issue, we propose a simple but effective energy self-normalization loss $\mathcal{L}_{al}^s$, which constrains the free energy of each classifier with a fixed anchor Δ, as Eq. 7.

$$\mathcal{L}_{al}^s = \mathbb{E}_{x\sim p_{data}^s}(\mathcal{F}_T^s(x) - \Delta)^2, \tag{7}$$

where Δ is a preset hyper-parameter, and the experimental results show that ESN is insensitive to its value. The total loss trained for every individual classifiers is given by Eq. 8.

$$\mathcal{L}_{total}^s = \mathbb{E}_{(x^s,y^s)\sim p_{data}^s}(\mathcal{L}_{ce}^s + \lambda\mathcal{L}_{al}^s), \tag{8}$$

where λ is a hyper-parameter to balance the $\mathcal{L}_{al}^s$ term. And we choose a representative temperature $T = 1$ during training. Our complete training algorithm is introduced in Alg. 1. Fig. 3 visualizes the free energy distribution with and without the self-normalization, which illustrates the effectiveness of the proposed ESN.

Voting with Stage-Cumulative Temperatures. As we have already normalized the Helmholtz free energy with the fixed anchor (Eq. 7), taking the negative Helmholtz free energy as the confidence score is a natural choice:

$$H^s(x) = -\mathcal{F}_T^s(x) = T \log \sum_{y \in \mathcal{Y}^s} \exp\left(h^s(x)[y]/T\right), \tag{9}$$

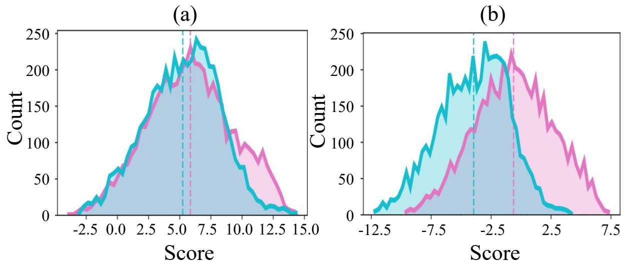

Figure 3: Distribution shift. We extract two stages' training data from Split-CIFAR100 and show their confidence scores trained with (a) and without (b) using our proposed anchor-based energy self-normalization. The y-axis is the count of images, and the x-axis is the confidence score.

Algorithm 1: Model Training

```
0: Given components: Pre-trained backbone $f_\theta$, stage classifier $f_\eta$, total stage number $S$, training iterations for each stage $M$, energy anchor $\Delta$, training data $\mathcal{D}$, learning rate $\epsilon$, temperature pool $\Omega$, candidate temperature pool $\Psi$;
for $s = 1, \cdots, S$ do
    Initialize classifier $f_{\eta_s}$ for the stage $s$;
    for $m = 1, \cdots, M$ do
        Draw a mini-batch of training data $B$ from $\mathcal{D}_s$;
        Calculate the logits $h^s(x) = f_{\eta_s}(f_\theta(x))$;
        Calculate the loss $\mathcal{L}_{total}^s$ by Eq.8;
        Update $\eta_s$ by $\eta_s \leftarrow \eta_s - \epsilon \bigtriangledown_{\eta_s} \mathcal{L}_{total}^s$;
    end for
    if $s > 1$ then
        for $t \in \Psi$ do
            Extract Helmholtz free energy $-\mathcal{F}_T^s(x)$ by Eq. 9;
            Calculate stage identification by $s^* = \mathrm{argmax}_{s\in\zeta}(-\mathcal{F}_T^s(x))$;
            Calculate stage identification accuracy $ACC_t$ of temperature $t$ by $\sum(s^* == s)$;
        end for
        $\Omega \leftarrow \mathrm{argmax}_{t\in\Psi}\, ACC_t$;
    end if
    Return the model parameters $\eta_s$.
end for
```

where $H^s(x)$ is the *logsumexp* of the logits with the control temperature parameter T. Previous energy-based out-of-distribution detection methods (Wang et al. 2022a; Liu et al. 2020) have shown that the in-distribution data usually has a lower free energy (i.e., a higher confidence score) than the out-of-distribution data for a certain classifier. Further aided by the energy self-normalization objective function, we can approximately derive that the right stage classifier for the in-stage data shall have the highest confidence score among all classifiers (Eq. 1). The derivation can be briefly expressed as $H^i(x^i) = H^j(x^j), H^j(x^j) > H^j(x^i) \rightarrow H^i(x^i) > H^j(x^i)$, where x^i is the in-stage data of the i-th stage but the out-stage data of the j-th stage. However, this derivation only approximately holds, we further propose a stage-

Algorithm 2: Inference

```
Given components: Pre-trained backbone $f_\theta$, stage classifiers $\{f_{\eta_s}\}_{s=1}^{S}$, temperature pool $\Omega$, total stage number $S$;
Input: Test example $x$;
Calculate image feature $l(x) = f_\theta(x)$;
for $s = 1, \cdots, S$ do
    Generate the $s$-th logits $h^s(x) = f_{\eta_s}(l(x))$;
    for $t \in \Omega$ do
        Calculate the scaled energy $-\mathcal{F}_T^s(x)$ by Eq. 9;
    end for
end for
Voting for stage identification $s^*$ by Eq.10;
Return the final prediction $y^*$ by Eq. 2.
```

cumulative temperature calibration strategy with a voting inference augmentation to further optimize the maximum confidence criterion (Eq. 1) without overfitting the newest stage's data.

As shown in Eq. (5), we can adjust the free energy by changing the temperature parameter T. Theoretically, we can find out the optimal temperature T for each classifier by optimizing the stage-ID prediction accuracy with all stages' data. This is however not possible in rehearsal-free incremental learning. As we can only access the current stage's data, we propose a stage-cumulative strategy to avoid overfitting.

Firstly, we find out the optimal temperature only with the current stage's training data by traversing the candidate temperatures Ψ and choosing the one with the best stage-ID prediction accuracy of the current stage. Secondly, we add this temperature to the final temperature pool denoted as Ω. Finally, we can traverse temperatures in the temperature pool Ω, and then aggregate stage-ID predictions under different temperatures by voting. For fairness and the comparability between different classifiers, we simultaneously change the temperature for all classifiers and do voting as Eq. 10.

$$s^* = \text{MODE}(\{\underset{s \in \zeta}{\text{argmax}} -\mathcal{F}_T^s(x) | \text{For } T \in \Omega\}), \qquad (10)$$

where $\text{MODE}(\cdot)$ is the mode operator to find the most frequent element in a collection. This voting-based inference augmentation strategy only increases negligible computation overhead. After the logits predicted by the model by once, we only need to recalculate Eq. (9) under different T.

Our augmented inference algorithm is introduced in Alg. 2, and the stage-cumulative temparature calibration is presented in Alg. 1.

Experiments

We conduct extensive experiments to evaluate the proposed ESN. Two traditional incremental learning scenarios are considered: (1) class-incremental learning where classes are generally from the same domain; (2) domain-incremental learning where classes are the same but from different domains. Moreover, we also consider a more general scenario, namely the cross-domain class incremental learning, where different classes are from diverse domains, and build a benchmark, named Split-DomainNet for this scenario. Consequently, four benchmark datasets are used for evaluation.

Split-DomainNet: We build the cross-domain class-incremental learning benchmark, Split-DomainNet, based on DomainNet (Peng et al. 2019). The Split-DomainNet is the scenerio where incoming data of each stage contains images of new categories from different domains. We construct this dataset as a challenging and practical scenario for the cross-domain class incremental learning. DomainNet collects images of 345 common objects from 6 diverse domains including Clipart, Real, Sketch, Infograph, Painting and Quickdraw. Because some domains and categories in DomainNet contain few instances (even without a single instance), we select the top 200 categories with the most images. We then split the 200 classes randomly into ten stages with 20 classes per stage. Instances of each stage come from a randomly selected domain.

Split-CIFAR100 (Wang et al. 2022c) splits the origin CIFAR-100 (Krizhevsky and Hinton 2009), which is a widely used benchmark dataset for class-incremental learning, into 10 sessions and each session has 10 classes.

5-Datasets (Ebrahimi et al. 2020) provides a benchmark for class incremental learning. Although each dataset in 5-Datasets is not difficult, it is still a challenging benchmark for pre-trained models, because there are slight similarity between them.

CORe50 (Lomonaco and Maltoni 2017) is a large benchmark dataset for continual object recognition. This dataset collects images of 50 different objects from 11 distinct domains (8 indoor and 3 outdoor). Three domains (3, 7, and 10) are selected as test set, and the remaining 8 domains are used for incremental learning. CORe50 is a benchmark for domain-incremental learning.

Evaluation Metrics. We use the *Final Average Accuracy* (FAA) and *Final Forgetting* (FF) as evaluation metrics for class-incremental learning and cross-domain task incremental learning, which are widely used in previous works (Mai et al. 2022). Moreover, we use the *Final Average Accuracy* for Domain-incremental learning.

Comparison Methods. We compare ESN against the state-of-the-art CIL and DIL methods. Though ESN is a rehearsal-free incremental learning method, we also consider rehearsal-based methods that need the buffer to store exemplars for a comparison. These methods include EWC (Kirkpatrick et al. 2017), LwF (Li and Hoiem 2017) ER (Chaudhry et al. 2019), GDumb (Prabhu, Torr, and Dokania 2020), BiC (Wu et al. 2019), DER++ (Buzzega et al. 2020) and Co2L (Cha, Lee, and Shin 2021), as well as the recently published transformer-based methods L2P (Wang et al. 2022c) and DyTox (Douillard et al. 2022). To compare fairly, we use the same ViT model pre-trained on ImageNet (i.e., ViT-B/16 (Dosovitskiy et al. 2021)) for all the competitors as well as ESN. We use the joint training result as the upper-bound for ESN on all the benchmarks.

Implementation details. We implement our method in PyTorch with two NVIDIA RTX 3090 GPUs. The proposed ESN is insensitive to hyper-parameters. We use the SGD optimizer and the cosine annealing learning rate scheduler

Method	Buffer size	FAA (↑)	FF (↓)
ER	1000	67.87±0.57	33.33±1.28
BiC		66.11±1.76	35.24±1.64
GDumb		67.14±0.37	-
DER++		61.06±0.87	39.87±0.99
Co^{2}L		72.15±1.32	28.55±1.56
DyTox		77.61±0.92	8.26±0.38
ER	5000	82.53±0.17	16.46±0.25
BiC		81.42±0.85	17.31±1.02
GDumb		81.67±0.02	-
DER++		83.94±0.34	14.55±0.73
Co^{2}L		82.49±0.89	17.48±1.80
DyTox		88.15±0.28	3.64±0.19
FT-seq	0	33.61±0.85	86.87±0.20
EWC		47.01±0.29	33.27±1.17
LwF		60.69±0.63	27.77±2.17
L2P		83.86±0.28	7.35±0.38
ESN		**86.34±0.52**	**4.76±0.14**
Upper-bound	-	91.27±0.18	-

Table 1: Results on Split-CIFAR100 for CIL. Bold: best rehearsal-free results. All results except ESN, DyTox, and Upper-bound are from (Wang et al. 2022c).

with a initial learning rate of 0.01 all benchmarks. We use 30 epochs for Split-CIFAR100 and Split-DomainNet, 10 epochs for 5-Datasets and CORe50. We set the batch size of 128 for all experiments. Momentum and weight decay parameters are set to 0.9 and 0.0005, respectively. We use ViT-B/16 (Dosovitskiy et al. 2021) pre-trained on ImageNet as backbone and the classifier is a class-attention block (CAB) (Touvron et al. 2021) with a linear projection. The hyper-parameters of CAB is the same as ViT-B/16 except the MLP ratio is 0.5, which has the parameters $3M$. Due to the fact raw features extracted from pre-trained ViT are not suitable for all downstream tasks, we also add parameters (10×768) to the input, like (Jia et al. 2022). The candidate temperature set Ψ is from a range of numbers from 0.001 to 1.0 with step of 0.001. We set the energy anchor $\Delta = -10$ and balance hyper-parameter $\lambda = 0.1$ for all benchmarks.

Comparison Results

We compare the proposed ESN with the state of the arts on Split-CIFAR100, Split-DomainNet, 5-Datasets and CORe50. We run ESN for 5 times randomly and report the average results. For fair comparisons, **all methods start from the same ImageNet pre-trained ViT-B/16.**

Results of Class-incremental learning. Table 1 and Table 2 summarize the results on Split-CIFAR100 and 5-Datasetsrespectively. ESN achieves state-of-the-art performance without any rehearsal buffer in terms of average accuracy and forgetting. We compute that ESN obtains a considerable relative improvement (an average of roughly 3.5%) over the best rehearsal-free methods. We can see that most rehearsal-based methods significantly improve by storing more data. This shows that the rehearsal-based methods' performances highly depend on buffer size. The outstanding performance of ESN indicates that the proposed anchor-based energy self-normalization can successfully aggregate all stage classifiers impartially, and thus can get outstanding performance even without rehearsal buffer.

Method	Buffer size	FAA (↑)	FF (↓)
ER	250	80.32±0.55	15.69±0.89
BiC		78.74±1.41	21.15±1.00
DER++		80.81±0.07	14.38±0.35
Co^{2}L		82.25±1.17	17.52±1.35
ER	500	84.26±0.84	12.85±0.62
BiC		85.53±2.06	10.27±1.32
DER++		84.88±0.57	10.46±1.02
Co^{2}L		86.05±1.03	12.28±1.44
FT-seq	0	20.12±0.42	94.63±0.68
EWC		50.93±0.09	34.94±0.07
LwF		47.91±0.33	38.01±0.28
L2P		81.14±0.93	4.64±0.52
ESN		**85.71±1.47**	**2.85±0.61**
Upper-bound	-	94.39±0.21	-

Table 2: Results on 5-Datasets for CIL. Bold: best rehearsal-free results. All results except ESN and Upper-bound are copied from (Wang et al. 2022c).

Method	Buffer size	FAA (↑)
ER	50/class	80.10±0.56
GDumb		74.92±0.25
BiC		79.28±0.30
DER++		79.70±0.44
Co^{2}L		79.75±0.84
DyTox		79.21±0.10
L2P		81.07±0.13
EWC	0	74.82±0.60
LwF		75.45±0.40
L2P		78.33±0.06
ESN		**91.80±0.31**
Upper-bound	-	92.50±0.11

Table 3: The final test accuracy on CORe50 for DIL. Bold: best rehearsal-free results. All results except ESN, DyTox, and Upper-bound are copied from (Wang et al. 2022c).

Results of Domain-incremental learning. Table 3 summarizes the results on the CORe50 dataset. CORe50 provides a challenging DIL benchmark that uses 8 domains as train set and 3 domains as test set. That means test images do not belong to any training domains, and this benchmark mainly tests the generalization ability after incremental learning. ESN achieves the best performance compared with other methods (about 17% improvements over L2P) with the same ViT-B/16 pre-trained backbone. Since there is no correct stage-ID for test images (no domain overlap), the accuracy of ESN comes from the ensemble voting strategy.

Results of Cross-Domain Class-incremental learning. Cross-Domain Class-incremental learning is a more challenging scenario than traditional CIL and DIL settings. As shown in Table 4, ESN out-performs all other rehearsal-free methods by a large margin (about 50% improvement). We can see that most class incremental learning algorithms fail to prevent catastrophic forgetting in the cross-domain setting to a great extent, as indicated by high final forget-

Method	Buffer size	FAA (↑)	FF (↓)
ER	250	64.54±1.06	28.21±0.45
BiC		66.99±1.27	19.91±0.23
DER++		70.18±0.37	21.31±0.55
DyTox		77.16±0.72	6.88±0.31
ER	500	70.90±1.35	21.49±0.61
BiC		68.19±1.22	21.76±0.39
DER++		74.61±0.27	16.65±0.94
DyTox		79.6±0.91	5.87±0.20
Finetune	0	35.66±2.73	59.89±2.05
EWC		22.35±1.86	76.11±1.28
LwF		28.86±1.92	64.91±1.01
L2P		45.65±0.82	15.26±0.51
ESN		**68.76±0.12**	**5.75±0.23**
Upper-bound	-	82.53±0.44	-

Table 4: Results on Split-DomainNet for cross-domain class-incremental learning. Bold: best rehearsal-free results.

Ablated components	FAA (↑)	FF (↓)
w/o energy self-normalization	80.21±1.93	9.35±0.59
w/o temperature calibration	85.73±2.04	4.88±0.70
w/o parameter isolation	83.94±1.80	6.42±0.45
None	86.34±0.52	4.76±0.14

Table 5: Ablation studies of the effect of related components. The experiments are performed on Split-CIFAR100.

ting (FF) shown in Table 4. Specially, some regularization-based methods, LwF and EWC, even perform worse than simply finetuning. This is probably due to some regularizations that are not robust to large domain shift. Our stage isolation learning strategy can preserve old knowledge successfully. Besides, the proposed anchor-based energy self-normalization strategy is robust to handle this challenging scenario.

Ablation Study

The effect of related components. To further study the effectiveness of ESN, we study the effect of our main components in Table 5. Table 5 (row 1) removes the proposed anchor-based energy self-normalization strategy $\mathcal{L}^s_{al}$, and keeps the other parts the same. The performance has a significant drop, suggesting that aligning all isolated classifiers to the same energy plane is the key issue in aggregating them impartially for final prediction. Table 5 (row 2) removes our proposed temperature selection strategy, and just using the default temperature 1 without voting for prediction. The results is slightly lower than ESN. The decrease suggests that using the proposed temperature calibration can further boost the performance. Table 5 (row 3) shares the same class-attention block (CAB) across tasks. As the result shows, parameter isolation is important in tackling catastrophic forgetting and maintaining performance.

The effect of different Δ. Δ is the main hyper-parameter of our proposed energy self-normalization loss, and we conduct an ablation study to investigate its effect. Table 6 shows the final results (FAA and FF) is insensitive to the value of Δ. That is probably because the most important thing is to normalize all classifiers to the same energy plane.

Δ	0	-3	-5	-15
FAA (↑)	85.96±1.07	85.59±0.42	86.20±0.19	86.25±0.18
FF (↓)	5.08±0.30	5.56±0.81	4.59±0.47	4.98±0.21

Table 6: The effect of the energy anchor Δ. The experiments are performed on Split-CIFAR100.

Method	Expansion		FAA (↑)	FF (↓)
	M	Inc (%)		
DER-ViT	86.6	100	83.43±1.87	5.52±0.64
DER-ResNet50	25.3	100	80.37±1.01	9.2±0.38
VPT	0.2	0.2	85.55±3.49	4.98±1.91
CAB	3.0	3.4	86.34±0.52	4.76±0.14

Table 7: Ablation studies of different network architectures. The experiments are performed on Split-CIFAR100.

The effect of different network architectures. In the main experiments, we mainly attach a class-attention block as a decoder to the pre-trained backbone. We point out that other network architectures can also use our proposed energy self-normalization method. Table 7 summarizes the results of using different architectures. Here, we add two parameter isolation methods to demonstrate our idea: VPT (Jia et al. 2022) and DER (Yan, Xie, and He 2021). VPT uses a small amount of task-specific learnable parameters into the input while freezing the other parts of the model to tune a pre-trained model to downstream tasks. DER expands a new network for each new coming task. The network can be any type, and we use both ResNet50 and ViT-B/16 for experiments. We report the amount of expansion parameters for a single incremental stage in the Table 7. Though the amount of expansion parameters of VPT is significantly less than CAB, VPT needs almost ten times inference time than CAB. That is because CAB works as a stage-specific decoder and uses a shared backbone to extract image features, which can decrease the computational expense. DER-like methods have the same inference speed problem and perform worse than VPT and CAB. The worse performance of DER-like methods is probably because training large models on a small subset of a dataset has severe over-fitting.

Conclusion

This paper proposes a novel rehearsal-free stage-isolation based general incremental learning framework. The proposed ESN learns stage-isolation classifiers for each stage, and uses then anchor-based energy self-normalization strategy to aggregate multiple isolated classifiers in an impartial way. Furthermore, we propose a control parameter (temperature) finding method and propose a voting based inference augmentation strategy for robust inference. Our experiments show that our method outperforms the current state-of-the-art on four large benchmarks by a large margin and can handle general incremental learning scenarios.

Acknowledgements

This work is funded by the National Key Research and Development Project of China (2019YFB1312000), the National Natural Science Foundation of China (62076195, 62206271, and U20B2052), the Fundamental Research Funds for the Central Universities (AUGA5710011522), and the Guangdong Basic and Applied Basic Research Foundation (2020B1515130004). This work is also supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (MSS21C002).

References

Al Rahhal, M. M.; Bazi, Y.; Al-Dayil, R.; Alwadei, B. M.; Ammour, N.; and Alajlan, N. 2022. Energy-based learning for open-set classification in remote sensing imagery. *International Journal of Remote Sensing*, 1–11.

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and Calderara, S. 2020. Dark experience for general continual learning: a strong, simple baseline. *NeurIPS*.

Cha, H.; Lee, J.; and Shin, J. 2021. Co2l: Contrastive continual learning. In *ICCV*.

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P. K.; Torr, P. H.; and Ranzato, M. 2019. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7): 3366–3385.

Dong, S.; Hong, X.; Tao, X.; Chang, X.; Wei, X.; and Gong, Y. 2021. Few-shot class-incremental learning via relation knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1255–1263.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

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

Du, Y.; and Mordatch, I. 2019. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32.

Ebrahimi, S.; Meier, F.; Calandra, R.; Darrell, T.; and Rohrbach, M. 2020. Adversarial continual learning. In *European Conference on Computer Vision*, 386–402. Springer.

Fort, S.; Ren, J.; and Lakshminarayanan, B. 2021. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34: 7068–7081.

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2019. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*.

Hendrycks, D.; Basart, S.; Mazeika, M.; Mostajabi, M.; Steinhardt, J.; and Song, D. 2019. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 831–839.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual Prompt Tuning. In *European Conference on Computer Vision (ECCV)*.

Joseph, K.; Khan, S.; Khan, F. S.; Anwer, R. M.; and Balasubramanian, V. N. 2022. Energy-based Latent Aligner for Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7452–7461.

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.

Knoblauch, J.; Husain, H.; and Diethe, T. 2020. Optimal continual learning has perfect memory and is np-hard. In *ICML*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; and Huang, F. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Li, X.; Zhou, Y.; Wu, T.; Socher, R.; and Xiong, C. 2019. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International Conference on Machine Learning*, 3925–3934. PMLR.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33: 21464–21475.

Liu, Y.; Hong, X.; Tao, X.; Dong, S.; Shi, J.; and Gong, Y. 2022. Model Behavior Preserving for Class-Incremental Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Lomonaco, V.; and Maltoni, D. 2017. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, 17–26. PMLR.

Mai, Z.; Li, R.; Jeong, J.; Quispe, D.; Kim, H.; and Sanner, S. 2022. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469: 28–51.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, 1406–1415.

Prabhu, A.; Torr, P. H.; and Dokania, P. K. 2020. Gdumb: A simple approach that questions our progress in continual learning. In *ECCV*.

Riemer, M.; Cases, I.; Ajemian, R.; Liu, M.; Rish, I.; Tu, Y.; and Tesauro, G. 2018. Learning to Learn without Forgetting by Maximizing Transfer and Minimizing Interference. In *ICLR*.

Serra, J.; Suris, D.; Miron, M.; and Karatzoglou, A. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, 4548–4557. PMLR.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30.

Tang, K.; Miao, D.; Peng, W.; Wu, J.; Shi, Y.; Gu, Z.; Tian, Z.; and Wang, W. 2021. CODEs: Chamfer Out-of-Distribution Examples against Overconfidence Issue. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1153–1162.

Tao, X.; Chang, X.; Hong, X.; Wei, X.; and Gong, Y. 2020a. Topology-preserving class-incremental learning. In *European Conference on Computer Vision*, 254–270. Springer.

Tao, X.; Hong, X.; Chang, X.; Dong, S.; Wei, X.; and Gong, Y. 2020b. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12183–12192.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 32–42.

Wang, H.; Li, Z.; Feng, L.; and Zhang, W. 2022a. ViM: Out-Of-Distribution with Virtual-logit Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4921–4930.

Wang, Y.; Huang, Z.; and Hong, X. 2022. S-Prompts Learning with Pre-trained Transformers: An Occam's Razor for Domain Incremental Learning. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; et al. 2022b. DualPrompt: Complementary Prompting for Rehearsal-free Continual Learning. *European Conference on Computer Vision*.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022c. Learning To Prompt for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 139–149.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large scale incremental learning. In *CVPR*.

Xie, J.; Lu, Y.; Zhu, S.-C.; and Wu, Y. 2016. A theory of generative convnet. In *International Conference on Machine Learning*, 2635–2644. PMLR.

Xie, J.; Yan, S.; and He, X. 2022. General Incremental Learning with Domain-aware Categorical Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14351–14360.

Xu, H.; and Zhang, J. 2020. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1959–1968.

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

Yang, D.; Zhou, Y.; Shi, W.; Wu, D.; and Wang, W. 2022a. RD-IOD: Two-Level Residual-Distillation-Based Triple-Network for Incremental Object Detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1): 1–23.

Yang, D.; Zhou, Y.; Zhang, A.; Sun, X.; Wu, D.; Wang, W.; and Ye, Q. 2022b. Multi-view correlation distillation for incremental object detection. *Pattern Recognition*, 131: 108863.

Zhang, X.; Dong, S.; Chen, J.; Tian, Q.; Gong, Y.; and Hong, X. 2022. Deep Class-Incremental Learning From Decentralized Data. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Zhao, B.; Chen, C.; Xiao, X.; Ju, Q.; and Xia, S. 2022. Energy Alignment for Bias Rectification in Class Incremental Learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3513–3517. IEEE.