Incremental Learning via Robust Parameter Posterior Fusion

ABSTRACT

The posterior estimation of parameters based on Bayesian theory is a crucial technique in Incremental Learning (IL). The estimated posterior is typically utilized to impose loss regularization, which aligns the current training model parameters with the previously learned posterior to mitigate catastrophic forgetting, a major challenge in IL. However, this additional loss regularization can also impose detriment to the model learning, preventing it from reaching the true global optimum. To overcome this limitation, this paper introduces a novel Bayesian IL framework, **R**obust **P**arameter **P**osterior **F**usion (RP^{2}F). Unlike traditional methods, RP^{2}F directly estimates the parameter posterior for new data without introducing extra loss regularization, which allows the model to accommodate new knowledge more sufficiently. It then fuses this new posterior with the existing ones based on the Maximum A Posteriori (MAP) principle, ensuring effective knowledge sharing across tasks. Furthermore, RP^{2}F incorporates a common parameter-robustness priori to facilitate a seamless integration during posterior fusion. Comprehensive experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets show that RP^{2}F not only effectively mitigates catastrophic forgetting but also achieves backward knowledge transfer.

CCS CONCEPTS

• **Computing methodologies → Lifelong machine learning**; *Bayesian network models*; Image representations.

KEYWORDS

Incremental Learning, Lifelong Learning, Catastrophic Forgetting, Bayesian Theory

ACM Reference Format:

. 2024. Incremental Learning via Robust Parameter Posterior Fusion. In *Proceedings of ACM MULTIMEDIA (MM).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Acquiring knowledge continuously is essential for intelligent systems operating in ever-changing environments, a process referred to as Incremental Learning (IL) [12, 33]. In contrast to the conventional training paradigm that amasses all data at once, IL adopts a task-by-task training paradigm, where only a subset of data is available for each training task. Training within this setting, models encounter the formidable challenge of catastrophic forgetting [17, 35], wherein the assimilation of new knowledge often results in diminished performance on previously learned tasks.

To mitigate catastrophic forgetting, a fundamental strategy is to identify and preserve the parameters critical to past tasks. This strategy has directed the development of numerous IL methodologies [1, 9, 34, 41, 68], including techniques grounded in Bayesian theory [20, 22, 32, 39, 46]. The Bayesian-based approaches utilize Laplace approximation to model the parameter posterior distribution of old tasks as a Gaussian distribution. This Gaussian posterior then serves as a priori to guide subsequent network training, enforcing regularization to prevent substantial deviations from this priori. In this context, the variance of the Gaussian distribution (i.e., the Hessian through Laplace approximation) acts as a parameter importance matrix to identify critical parameters.

While the regularization has proven its effectiveness in IL, controlling its strength poses a considerable challenge for Bayesian-based methods. A particular issue is ensuring that parameter adjustments through regularization accurately reflect the Maximum A Posteriori (MAP) estimate of the joint posterior distribution for both new and old tasks. Particularly, achieving a harmonious integration of the regularization (representing the old-task posterior) with the learning loss for new tasks remains a challenge [21], even when employing predefined weights. Furthermore, the necessity for accommodating new data within IL contexts inevitably compels the parameters to diverge from the old-task posterior, potentially leading to information loss.

To address the above challenges, this paper introduces a novel Bayesian IL framework, referred to as Robust Parameter Posterior Fusion (RP^{2}F). Unlike traditional methods that integrate the posterior of old tasks through regularization, RP^{2}F constructs a parameter posterior for the new task and then directly fuses it with the posterior from previous tasks. This integrated posterior enables the derivation of a closed-form MAP solution for parameters, facilitating a more coherent balance across tasks. As the learning process evolves, parameter estimates continuously adjust from the old-task MAP with the introduction of each new task, which may trigger potential forgetting. To mitigate this, RP^{2}F incorporates a parameter robustness priori within the training regimen, which bolsters the model's resilience against changes in parameters.

To implement RP^{2}F, two critical techniques are imperative for enhancing performance: the computation of the Hessian matrix in Laplace approximation and the design of the parameter robustness priori. Existing methods [22, 46] typically approximate the diagonal of the Hessian matrix with the Fisher information matrix due to the computational and storage constraints, where the Fisher information matrix only depends on the model's first-order derivative information. Instead, this paper develops a perturbation-based method for estimating the Hessian information, which leverages derivatives with respect to small perturbations applied to the parameters. Furthermore, our analysis reveals that a network's robustness is closely linked to the uniformity of the singular values of the extracted feature matrix. Drawing on this insight, we propose regularizing the features during training to balance the singular values, thus diminishing the potential forgetting issue caused by parameter changes.

MM, 2024, Melbourne, Australia

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

In summary, the core contributions of this research are listed as follows:

- We propose a novel Robust Posterior Parameter Fusion (RP^{2}F) framework for IL, which leverages the Bayesian theory to derive a posterior fusion method across tasks.
- To facilitate Laplace posterior estimation, we develop a Hessian approximation method via parameter perturbation. This method captures second-order derivative information while maintaining the computational complexity of first-order gradients.
- We demonstrate that the uniformity of singular values in the extracted features is critical for model robustness. Based on this insight, we introduce a parameter robustness priori for RP^{2}F to achieve a seamless posterior fusion.
- We conduct several experiments on the CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets, demonstrating that RP^{2}F supports backward knowledge transfer and achieves state-of-the-art performance.

2 RELATED WORKS

2.1 Incremental Learning

Incremental learning focuses on developing algorithms that can continuously learn from new data without forgetting previously acquired knowledge [12, 33]. Approaches within this field are broadly categorized into regularization-based methods, rehearsal-based methods, and dynamic architecture methods. In the following, we will review these three class methods.

Regularization-based methods integrate regularization loss terms to prevent forgetting. Some methods [1, 22, 29, 68] punish the changes in critical parameters containing old-task knowledge. Other knowledge-distillation-based methods hinge on the principle of transferring knowledge from an older model (teacher) to a newer model (student) during the IL process. A seminal work is Learning without Forgetting (LWF) proposed by Li and Hoiem [28], which utilizes knowledge distillation to preserve previously learned information while accommodating new knowledge. Based on LWF, several other notable methods [14, 26, 27, 47, 56, 57, 69, 71] have been developed. **Rehearsal-based IL methods** operate by retaining a subset of the original training data [3, 5, 6, 8, 9, 30, 38, 42] or generating synthetic data [49, 50, 65], which is then replayed with new data during subsequent training phases. Dark Experience Replay (DER) [7] combines experience replay with a distillation mechanism to maintain a balance between old and new knowledge. DER selectively stores a subset of the data from previous tasks and rehearses it alongside new data, thus preserving the model's performance on historical tasks. On the other hand, generative-replay-based methods leverage generative models, such as Generative Adversarial Networks (GANs) [16], to synthesize data from previous tasks, thereby eliminating the need to store real data. **Dynamic architecture methods** typically involve modifying the architecture of neural networks dynamically in response to new tasks. Some methods are based on masks [10, 19, 23, 34], such as PackNet [34]. PackNet prunes weights that are non-essential for learned tasks, thereby freeing up network capacity for new tasks, a process that strategically balances between retaining learned knowledge and adapting to new information. On the other hand, expansion-based methods [40, 53–55, 64] like Dynamically Expandable Networks (DEN) [66] selectively expand the network by adding neurons as needed.

As a subset of regularization-based methods, bayesian-based IL methods [20, 22, 25, 37, 39, 46, 61] utilise Bayesian inference through the learning process. Elastic Weight Consolidation (EWC) [22] applies a Bayesian perspective to determine the importance of neural network parameters, followed by imposing a constraint on significant deviations from previous parameters. IMM [25] proposes to merge the neural network parameters from previous tasks with the current task, which shares a similar motivation with the posterior fusion of RP^{2}F. Beyond IMM, RP^{2}F further introduces parameter robustness priori and a more precise perturbation-based Hessian estimation method.

2.2 Robust Model Training

In the field of robust model training, the perturbation-based method is a primary research direction within the machine learning community. A common practice involves modeling parameter robustness using Lipschitz continuity and enhancing robustness by constraining the upper bound of the Lipschitz constant [2, 11, 18, 58]. However, recent studies [43] have highlighted that the Lipschitz constant is sensitive to the scale of model parameters and input features, and propose to employ condition numbers as a more precise metric of robustness. In the context of IL, robust methodologies have been integrated into some exemplar-based methods. They primarily concentrate on model robustness to exemplar variations. For instance, LiDER [5] forces the model to become robust to the changes of exemplars, thereby mitigating the over-fitting issue in exemplar-based methods. DRO [63] continually evolves the exemplar buffer to ensure the model learns more robust features. In contrast to these methods, our proposed RP^{2}F does not rely on exemplar samples. Instead, we focus on investigating robustness against parameter variations, with the goal of mitigating information losses in posterior fusion, which contributes a novel perspective to IL.

3 PRELIMINARY

3.1 Problem Definition

Incremental learning (IL) is a cognitive process characterized by the gradual acquisition and assimilation of knowledge over time. As a popular setting of IL, Task Incremental Learning (Task-IL) [59] focuses on the sequential acquisition and retention of knowledge across multiple related tasks, where each task may involve distinct but related learning objectives.

Formally, let $D_1, D_2, \ldots, D_T$ denote datasets of T classification tasks, where $D_t = \{(x_k^t, y_k^t \in Y_t)\}_{k=1}^{n_t}$ is associated with task t that contains n_t input-label pairs. These tasks feature disjoint classes, i.e., $Y_{t_1} \cap Y_{t_2} = \emptyset, \forall t_1 \neq t_2$. Task-IL sequentially presents these tasks to an IL model, with only the associated dataset D_t being available at time step t. The model must adapt to task t while also retaining knowledge of the former tasks. During the testing phase, the model's overall performance across all encountered tasks will be evaluated.

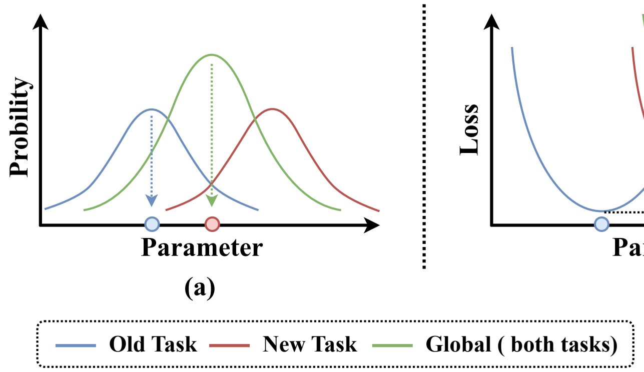


Figure 1: (a) RP^{2}F identifies the optimal parameter at the peak of the fusion posterior distribution. (b) After learning a new task, parameter adjustments shift from the peak posterior of the old task toward the maximum fusion posterior, leading to increased loss on the old task, i.e., forgetting. (c) The forgetting can be mitigated by enhancing parameter robustness against substantial changes.

3.2 Bayesian Incremental Learning

Incremental learning has been extensively explored within the Bayesian framework. The Bayesian IL framework leverages Bayes' rule to systematically update the posterior parameter distribution as new data becomes available. The fundamental concept lies in the Bayes' theorem, which provides an estimation of the parameter posterior as follows:

$$P(\theta|D) \propto P(D|\theta)P(\theta), \tag{1}$$

where $P(\theta|D)$ represents the parameter posterior after observing data D. This posterior is shaped by the parameter priori $P(\theta)$ and the likelihood $P(D|\theta)$.

In the context of IL, the posterior estimation after observing t tasks in Eq. 1 can be extended into multiple tasks [20] as follows:

$$P(\theta|D_1,\ldots,D_t) \propto P(\theta)\prod_{t'=1}^{t} P(D_{t'}|\theta). \tag{2}$$

Using the posterior of all the former tasks as a priori, the posterior can be computed cumulatively:

$$P(\theta|D_1,\ldots,D_t) \propto P(\theta|D_1,\ldots,D_{t-1})P(D_t|\theta). \tag{3}$$

An integral part of Bayesian-based IL methods is the estimation of the posterior distributions $P(\theta|D_1,\ldots,D_{t-1})$, which is often computationally challenging for complex models. The Laplace approximation [31] offers a tractable solution by approximating the posterior distribution with a Gaussian centred at the maximum a posteriori (MAP) estimate. This approximation is particularly useful when the posterior is unimodal. The mathematical representation of this approximation is formalized as follows:

$$P(\theta|D) \approx \mathcal{N}(\theta|\mu_{MAP}, H_{MAP}^{-1}), \tag{4}$$

where μ_{MAP} is the mode of the posterior distribution (i.e., the MAP estimation), and H_{MAP}^{-1} is the inverse of the Hessian around μ_{MAP}, providing a second-order approximation of the posterior curvature.

Building upon this framework, bayesian-based IL methods [20, 22, 25] emerge to incorporate Bayesian principles into neural network training for each task. They typically mitigate catastrophic forgetting by introducing a regularization term to the loss function that penalizes changes to important parameters:

$$\begin{aligned}\theta_{1:t} &= \arg\max_{\theta} \log P(\theta|D_1,\ldots,D_t)\\ &= \arg\max_{\theta} \log P(D_t|\theta)+\log P(\theta|D_1,\ldots,D_{t-1})\\ &\approx \arg\min_{\theta} L_{ce}(\theta,D_t)+\frac{\lambda}{2}(\theta-\mu_{1:t-1})^{\top}\Lambda_{1:t-1}(\theta-\mu_{1:t-1}),\end{aligned} \tag{5}$$

where L_{ce} represents the cross-entropy loss and λ stands for a pre-defined hyperparameter weighting the regularizer. Here, the posterior $P(\theta|D_1,\ldots,D_{t-1})$ is approximated by the Gaussian distribution $\mathcal{N}(\theta|\mu_{1:t-1},\Lambda_{1:t-1})$, where $\mu_{1:t-1}$ is typically assigned to the value of $\theta_{1:t-1}$ and $\Lambda_{1:t-1}$ signifies the importance matrix. The continuous update of $\Lambda_{1:t-1}$ incorporates the Hessian associated with the new task [22]:

$$\Lambda_{1:t} = \Lambda_{1:t-1} + H_t = \Lambda_{1:t-1} + \nabla^2_{\theta} L_{ce}(\theta,D_t)|_{\theta=\theta_{1:t}}. \tag{6}$$

However, the computation of the Hessian matrix poses a significant challenge, due to its computational complexity (second-order partial derivatives) and the extensive storage requirements. In practice, a prevalent approach is to approximate the Hessian using the Fisher information matrix [20, 22, 39, 46]:

$$H_t \approx F_t = \nabla_{\theta} L_{ce}(\theta,D_t)\nabla_{\theta} L_{ce}(\theta,D_t)^{\top}|_{\theta=\theta_{1:t}}. \tag{7}$$

4 METHOD

4.1 Robust Bayesian IL Framework

Although Bayesian-based IL methods have demonstrated effectiveness in mitigating forgetting, they encounter significant challenges. One such challenge is **the control of regularization strength**. In Eq. (5), the cross-entropy loss and the regularization terms may not be scaled equivalently. This imbalance may lead $\theta_{1:t}$ fails to accurately represent the MAP estimation of posterior $P(\theta|D_1,\ldots,D_t)$. Manually adjusted balancing coefficients can not solve this issue, especially in the complex and unpredictable trajectory of neural network training. Additionally, **the issue of forgetting** arises as parameters transition from the local optimum associated with previous tasks $\theta_{1:t-1}$ to a global optimum that encompasses new learning objectives $\theta_{1:t}$ (as illustrated in Figure 1 (b)). This forgetting becomes noteworthy when the performance is sensitive to parameter changes.

To address the aforementioned challenges, we propose the Robust Parameter Posterior Fusion (RP^{2}F) framework for IL. Specifically, RP^{2}F reformulates Eq. (2) by choosing a specific parameter robust priori $P(\theta)$ as follows:

$$P(\theta|D_1,\dots,D_t) \propto \prod_{t'=1}^{t} P(\theta)^{\frac{1}{t}} P(D_{t'}|\theta). \tag{8}$$

Equation (8) enables us to model task-specific posterior solely based on the data from each respective task. In other words, instead of modeling the posterior of all encountered tasks $P(\theta|D_1,\dots,D_{t-1})$ in Eq. (3), we are able to yield the MAP estimation directly for each individual task:

$$\begin{aligned}\theta_t^* &= \arg\max_\theta \ \log P(\theta)^{\frac{1}{t}} P(D_t|\theta) \\ &= \arg\max_\theta \ \log P(D_t|\theta) + \frac{1}{t}\log P(\theta) \\ &\approx \arg\min_\theta L_{ce}(\theta, D_t) + \lambda L_{\text{robust\_pri}}(\theta),\end{aligned} \tag{9}$$

where $L_{\text{robust\_pri}}(.)$ represents a parameter robustness priori weighted by λ. Notably, different from Eq. (5), Eq. (9) derives the MAP estimation only for the single task t.

Based on θ_t^*, we apply the Laplace approximation to estimate $P(\theta)^{\frac{1}{t}} P(D_{t'}|\theta)$ with Gaussian $\mathcal{N}(\theta|\theta_t^*, H_t^{*-1})$, where H_t^* represents the Hessian of the error around θ_t^*, i.e., $H_t^* = \nabla_\theta^2 L_{ce}(\theta, D_t)|_{\theta=\theta_t^*}$. Taking into account the practical constraints of storage and computation, our approach aligns with prior works [22] that employ a diagonal matrix to estimate the Hessian matrix. The diagonal matrix, denoted by Λ_t^*, is thus regarded as indicative of the importance of the parameters. Building upon the Laplace approximation, we are able to derive the MAP estimation $\theta_{1:t}^*$ of the posterior distribution in Eq. (8):

$$\begin{aligned}\theta_{1:t}^* &= \arg\max_\theta P(\theta|D_1,\dots,D_t) \\ &\approx \frac{\sum_{t'=1}^{t} \Lambda_{t'}^* \theta_{t'}^*}{\sum_{t'=1}^{t} \Lambda_{t'}^*}.\end{aligned} \tag{10}$$

For the proof of Eq. (10), please refer to the supplementary materials.

Additionally, we present a task-accumulative version, designed to fulfil the requirements of IL settings:

$$\theta_{1:t}^* \approx \frac{\Lambda_{1:t-1}^* \theta_{1:t-1}^* + \Lambda_t^* \theta_t^*}{\Lambda_{1:t-1}^* + \Lambda_t^*}, \tag{11}$$

where $\Lambda_{1:t}^* = \Lambda_{1:t-1}^* + \Lambda_t^*$ also update in an accumulative manner.

Compared to the conventional Bayesian-based IL methods, RP^{2}F has two specific designs. Firstly, RP^{2}F determines the optimal parameter through the MAP estimate of the fusion parameter posterior for all encountered tasks in Eq (10). Unlike regularizer-based methods that struggle to balance the regularizer in Eq. (5), the posterior fusion manner facilitates a harmonious knowledge accumulation, achieving an enhanced equilibrium between maintaining previously learned information (stability) and incorporating new knowledge (plasticity). Secondly, the incorporation of the parameter-robustness priori mitigates potential forgetting induced by parameter updating.

To implement the RP^{2}F framework, two key components remain to be devised: the efficient estimation of Hessian and the design of a parameters-robustness priori. Conventionally, the Hessian is estimated using the Fisher information matrix. Although the Fisher information matrix possesses many desirable properties [22], as described in Eq. (7), the common computation of Fisher inherently contains only first-order gradient information. This limitation is especially highlighted as previous works [22, 46] consider only the diagonal elements of the Fisher information matrix. To address this limitation, this paper introduces a novel Hessian matrix approximation method based on parameter perturbation, which effectively approximates second-order derivative information with the same computational complexity as first-order derivatives. In terms of the robustness priori to parameters, our investigation reveals that the uniformity in the eigenvalues of the features extracted by the network is closely linked to the robustness of the parameters. This observation motivates us to average the eigenvalues of features during training.

Next, Sections 4.2 and 4.3 elaborate on the Hessian approximation method and the parameter robustness priori, respectively.

4.2 Hessian Approximating via Parameter Perturbation

Previous works [22, 46] typically employ the diagonal Fisher information matrix to estimate the Hessian, focusing exclusively on the first-order derivatives as outlined in Eq. (7). In contrast, our research presents a novel Hessian estimation method based on parameter perturbation. The estimation is based on the gradient corresponding to infinitesimal parameter shifts, thereby providing a more precise and efficient approximation of Hessian.

To be more specific, considering the MAP estimation for t-th task posterior θ_t^*, we introduce a minor perturbation δ to θ_t^*. The second-order Taylor expansion on the empirical risk function is formulated as follows:

$$L_{ce}(\theta_t^* + \delta, D_t) \approx L_{ce}(\theta_t^*, D_t) + \nabla L_{ce}(\theta_t^*, D_t)^\top \delta + \frac{1}{2}\delta^\top H_t^* \delta, \tag{12}$$

where $H_t^* = \nabla_\theta^2 L_{ce}(\theta, D_t)|_{\theta=\theta_t^*}$ denotes the Hessian around θ_t^*.

The expansion facilitates the analytical derivation of the model's sensitivity to parameter changes. Considering that θ_t^* is sufficiently trained such that the first-order gradient approaches zero, by setting $\delta = \eta\mathbf{1}$ where $\mathbf{1}$ signifies an all-one vector and η represents an infinitesimally small constant, we differentiate both sides of Eq. (12) with respect to the perturbation δ yields the sum rows of the Hessian matrix:

$$\eta H_t^* \mathbf{1} \approx \left.\frac{\partial L_{ce}(\theta_t^* + \delta, D_t)}{\partial \delta}\right|_{\delta=\eta\mathbf{1}}. \tag{13}$$

By taking the assumption that the Hessian of a single-layer approaches diagonalization [44, 45], we estimate all the diagonal entries of H_t^* with Eq. 13 layer by layer, ultimately forming the diagonal matrix Λ_t^*.

Estimating Hessian via perturbation significantly simplifies the complexity of Hessian computation. For a network with n parameters, this method boasts a computational complexity of $O(n)$. In contrast, the direct computation of the Hessian, considering its need to evaluate second-order derivatives for every pair of parameters, generally exhibits a computational complexity of $O(n^2)$, or even $O(n^3)$ when matrix inversion is involved. The efficient Hessian estimation enables a scalable and high-performance implementation of RP^{2}F.

4.3 Enhancing Parameter Robustness via Feature Regularization

In Section 4.1, by maximizing the fusion posterior, we derive the optimal estimation $\theta^*_{1:t}$. Intuitively, we aspire for $\theta^*_{1:t}$ to exhibit linear connectivity with the local optima for each task (e.g., θ^*_t), aiming to minimize the difference $f(x;\theta^*_t) - f(x;\theta^*_{1:t})$ for any $x \in D_t$. In other words, as depicted in Figure 1, our objective is to ensure that θ^*_t is robust to parameter changes, which motivates us to incorporate robustness priori into parameters during training.

The robustness of parameters is commonly characterized by the Lipschitz constant. However, the Lipschitz constant is sensitive to the scale of both the data and model parameters. As an alternative, inspired by [43], our work introduces a relative metric to assess the robustness of the model, which compares the rate of change in model outputs against the rate of parameter alterations:

$$R(\theta) = \sup_{\delta} \sum_{x\in D} \frac{\|f(x;\theta+\delta) - f(\theta)\|/\|f(x;\theta)\|}{\|\delta\|/\|\theta\|}, \tag{14}$$

where δ represents a perturbation applied to θ. The numerator in (14) captures the relative change in model output due to parameter perturbation, while the denominator reflects the relative scale of the parameter perturbation. This formulation thus provides a measure of output sensitivity that is normalized by both the extent of the perturbation and the scale of the parameters, offering a more precise evaluation of the model's robustness.

Before delving into the robustness of deep neural models, we begin our examination with a simpler linear model $f(X;\Theta) = \Theta X$, where X denotes the input feature matrix with each column corresponding to an individual input sample $x \in D$. For this linear model, the relative-robustness metric is bounded by:

$$\begin{aligned} R(\Theta) &= \sup_{\Delta} \frac{\|(\Theta+\Delta)X - \Theta X\|/\|\Theta X\|}{\|\Delta\|/\|\Theta\|} \\ &= \sup_{\Delta} \frac{\|\Delta X\|}{\|\Delta\|}\frac{\|\Theta\|}{\|\Theta X\|} = \|X\|\frac{\|\Theta\|}{\|\Theta X\|} \leq \|X\|\|X^{\dagger}\|, \end{aligned} \tag{15}$$

where $X^{\dagger}$ represents the pseudo-inverse of X. When employing the Euclidean norm (also known as the ℓ_2 norm) for the metric, $R(\Theta)$ is bounded by the ratio between the maximum singular value of X to its minimum singular value:

$$R(\Theta) \leq \frac{s_{max}(X)}{s_{min}(X)}. \tag{16}$$

Building upon the robustness analysis in the linear scenario, we extend our investigation to a deep neural network composed of L layers [20]:

$$R(\theta) \leq \prod_{l=1}^{L} R(\sigma^l) \prod_{l=1}^{L} R(\Theta^l) \leq M \prod_{l=1}^{L} \frac{s_{max}(X^l)}{s_{min}(X^l)}, \tag{17}$$

where X^l, Θ^l, and σ^l denote the input features, parameters, and activation units of the l-th layer, respectively. We suppose that the robustness of all activation units is bounded by a positive constant M, i.e., $\prod_{l=1}^{L} R(\sigma^l) \leq M$.

Equation (17) elucidates that the robustness of the network parameters is governed by the uniformity of the singular values of the input features at each layer (extracted by the preceding layer). Yujun et al. [48] have demonstrated that this uniformity can be enhanced by penalizing the Frobenius norm of the covariance matrix. Based on this insight, we introduce the following priori to enhance the robustness of network parameters:

$$L_{\text{robust\_pri}}(\theta) = \frac{1}{L}\sum_{l=1}^{L} \|\hat{X^l}^{\top}\hat{X^l}\|_F^2, \tag{18}$$

where $\hat{X^l}$ represents the normalized feature matrix of the l-th layer. However, Eq. (18) presents a considerable challenge in terms of computational complexity, particularly for high-dimensional deep neural networks. Given that features in the shallow layers of neural networks tend to approach a full rank [15], in practice, we strategically apply regularization only to the features extracted from the penultimate layer.

4.4 Incremental Learning with Robust Parameter Posterior Fusion (RP^{2}F)

This section elaborates on our IL algorithm, which leverages the robust parameter posterior fusion framework. This approach integrates the perturbation-based Hessian approximation and the parameter-robustness priori proposed in Section 4.2 and 4.3, respectively. We will describe the training procedure of our method on a multi-head model consisting of a feature extractor θ and T task-specific classifier $\tau_1, \ldots, \tau_T$.

Learning task 1. During the phase of Task 1, the feature extractor alongside the classifier undergoes joint training using dataset D_1. The training leverages cross-entropy loss and parameter robustness priori to enhance performance and ensure robustness:

$$\theta^*_1, \tau^*_1 \longleftarrow \operatorname*{arg\,min}_{\theta_1,\tau_1} L_{ce}(\tau_1 \circ \theta_1, D_1) + \lambda L_{\text{robust\_pri}}(\tau_1 \circ \theta_1, D_1), \tag{19}$$

where $\tau \circ \theta$ denotes a unified network that the classifier τ operating on features extracted by the feature extractor θ. Upon the completion of training for task 1, we assign the parameters of an additional fused posterior extractor with $\theta^*_{1:1} \longleftarrow \theta^*_1$, and approximate the Hessian via $\Lambda^*_{1:1} \longleftarrow \frac{\partial L_{ce}(\theta^*_1+\delta, D_t)}{\eta\partial\delta}$.

Learning the following tasks (t for illustration). We initialize θ_t with the last optimal parameters θ^*_{t-1} learned from the previous task. This strategy aims to keep the optimization process for all tasks within the same vicinity of a local optimum, thereby facilitating effective fusion of the posterior distribution [25, 52]. Subsequently, during **each training epoch**, the model undergoes a four-step updating process: 1) training the feature extractor θ_t; 2) estimating the Hessian Λ_t surrounding θ_t; 3) updating the fused posterior extractor $\theta_1 : t$; 4) training the task-specific classifier τ_t.

The formulations for these four steps are presented as follows:

$$\begin{aligned}
\theta_t &\longleftarrow \theta_t - \alpha \nabla_{\theta_t} L_{ce}(\tau_t \circ \theta_t, D_1) + \lambda L_{\text{robust\_pri}}(\tau_t \circ \theta_t, D_t) \\
\Lambda_t &\longleftarrow \frac{\partial L_{ce}(\theta_t + \delta, D_t)}{\eta \partial \delta} \\
\theta_{1:t} &\longleftarrow \frac{\Lambda^*_{1:t-1}\theta^*_{1:t-1} + \Lambda_t \theta_t}{\Lambda^*_{1:t-1} + \Lambda_t} \\
\tau_t &\longleftarrow \tau_t - \alpha \nabla_{\tau_t} L_{ce}(\tau_t \circ \theta_{1:t}, D_1),
\end{aligned} \tag{20}$$

where α denotes the learning rate. This iterative update is designed to ensure that the classifier τ_t remains compatible with the optimal fused posterior extractor $\theta_{1:t}$. Upon the training convergence, we obtain the optimal values of θ^*_t, Λ^*_t, $\theta^*_{1:t}$, and τ^*_t. Among them, Λ^*_t is utilized to refine the variance of the fused posterior $\Lambda^*_{1:t} \longleftarrow \Lambda^*_{1:t-1} + \Lambda^*_t$. $\Lambda^*_{1:t}$, along with the others, is maintained for the subsequent task training.

Inference. During the testing phase, given sample x associated with task identification t, we select the corresponding classifier τ^*_t along with fused extractor $\theta^*_{1:T}$ to perform a prediction:

$$\hat{y} = f(x; \tau^*_t \circ \theta^*_{1:T}), \tag{21}$$

where $\hat{y}$ denotes the classification result.

5 EXPERIMENTS

5.1 Settings

Datasets. To ensure a comprehensive evaluation, we select the following three datasets to conduct experiments:

- **5-split CIFAR-10 [24]**. The CIFAR-10 dataset consists of 60,000 color images of 32 × 32 pixels distributed across 10 different classes, with each class containing 6,000 images. In our experiments, CIFAR-10 is divided into 5 splits based on the class labels.
- **10-split CIFAR-100 [24]**. Similar to CIFAR-10, CIFAR-100 features 100 classes containing 600 images each. We partition CIFAR-100 into 10 splits, with each containing 10 classes.
- **10-split Tiny-ImageNet [51]**. Tiny-ImageNet is a scaled-down version of the ImageNet dataset, consisting of 200 classes, each with 600 64 × 64 color images. For our study, this dataset is segmented into 10 splits, each with 20 classes.

Performance metrics. To thoroughly evaluate the performance of RP^{2}F, we employ two metrics [7, 30]: classification accuracy (ACC) and backward knowledge transfer (BWT). ACC is calculated after the model has been sequentially trained on all tasks. This metric serves as a direct indicator of the model's ability to classify images correctly across all classes, reflecting its overall performance in IL. BWT is designed to quantify how learning new tasks affects the performance of previously learned tasks. A negative value of BWT indicates forgetting, while a positive value suggests that learning new tasks has beneficial effects on previous tasks' performance. The formula for calculating BWT is defined as follows:

$$BWT = \frac{1}{T-1} \sum_{t'=1}^{T-1} (ACC_{T,t'} - ACC_{t',t'}), \tag{22}$$

where $ACC_{j,i}$ represents the accuracy on task i once the model finish learning task j. To ensure the reliability of results, we conduct all experiments five times with random seeds and report the average performance.

Baselines. We compare our method with various latest and classic IL methods, including Learning without Forgetting (LwF) [28], Synaptic Intelligence (SI) [68], Gradient Episodic Memory (GEM) [30], online Elastic Weight Consolidation (oEWC) [46], Learning without Memorizing (LwM) [13], Dark Experience Replay (DER and DER++) [7], Efficient Feature Transformation (EFT) [60], Pototype Augmentation and Self-Supervision (PASS) [70], Gradient Projection Memory (GPM) [41], Adam-NSCL [62], Always Be Dreaming (ABD) [50], Complementary Learning System (CLS-ER) [3], Filter Atom Swapping (FAS) [36], DCPOC [53], PRAKA [47], and MIND [4]. We also report the performance of a base model (referred to as Joint), which is trained jointly using data from all tasks. Clearly, Joint does adhere to the Task-IL setting, and its results are typically considered as the upper bound for IL methods. Additionally, for methods specifically designed for Class-IL, we employ a multi-head version of them to obtain the experimental results.

Implementation details. The experiments are conducted following the requirements of the Task-IL setting, where only the corresponding dataset is available for training in each task. For exemplar-based methods, we provide them with an extra sample buffer that can store up to 500 samples. We optimize the model parameters using Stochastic Gradient Descent (SGD), with learning rates adjusted specifically for each dataset: 0.2 for CIFAR-10, 0.05 for CIFAR-100, and 0.3 for Tiny-ImageNet. The hyperparameter λ (weighting the parameter-robustness regularize in Eq. (9)) is set to 1e-6 for CIFAR-10 and 1e-5 for both CIFAR-100 and Tiny-ImageNet. Further details are available in our supplementary code.

5.2 Comparison Result

In this section, we present a comprehensive comparison of our proposed RP^{2}F method against several state-of-the-art baselines as detailed in Section 5.1. Table 1 presents the average accuracy (%) over five runs on 5-split CIFAR-10, 10-split CIFAR-100, and 10-split Tiny-ImageNet.

The Joint model, which serves as the theoretical upper bound performance, consistently shows the highest accuracy across all datasets. We also found that some replay-based methods, such as GEM, DER, DER++, and ABD, exhibit competitive performance on small-scaled CIFAR-10, but experience a decline in efficacy on the CIFAR-100 and Tiny-ImageNet datasets. This observation suggests that the replay buffers and generative models may be insufficient for modeling complex data distributions.

Additionally, EFT achieves the highest accuracy on the CIFAR-10 dataset with 95.09%, surpassing RP^{2}F. The observed discrepancy in the performance of RP^{2}F may be attributed to the inherent structure of CIFAR-10's 5-split configuration, where each task only contains data of two classes. This limited class diversity may lead to overfitted parameters, which further affect the accurate estimation of posterior within RP^{2}F. Nevertheless, RP^{2}F outperforms all other baselines on the more challenging CIFAR-100 and Tiny-ImageNet datasets, suggesting its superior capability to handle more diverse and complex datasets.

Table 1: Comparison results on several datasets. We report the average accuracy (%) over five runs with random seeds, and the higher the better. (*) indicates the upper-bound model that is jointly trained with all tasks.

Methods	Venue	CIFAR-10	CIFAR-100	Tiny-ImageNet	Average
Joint*	-	98.07	91.18	82.01	90.42
LwF [28]	TPAMI2017	91.91±0.7	63.78±4.3	58.61±1.8	71.43
SI [68]	ICML2017	76.15±2.6	62.21±2.6	60.91±1.3	66.42
GEM [30]	NIPS2017	85.14±2.1	62.80±2.7	44.66±1.7	64.20
oEWC [46]	ICML2018	64.17±4.8	38.40±1.9	31.91±0.9	44.83
LwM [13]	CVPR2019	78.01±0.8	68.88±0.9	45.57±0.2	64.15
DER [7]	NIPS2020	93.13±0.3	73.26±1.3	51.22±1.5	72.54
DER++ [7]	NIPS2020	93.71±0.4	74.86±1.1	53.00±0.4	73.86
EFT [60]	CVPR2021	95.09±0.3	79.29±0.6	63.88±0.5	79.42
PASS [70]	CVPR2021	86.07±0.2	77.30±0.4	62.87±0.4	75.41
GPM [41]	ICLR2021	86.58±0.9	70.93±0.9	59.84±0.2	72.45
Adam-NSCL [62]	CVPR2021	87.23±0.4	65.69±0.2	59.98±0.7	70.97
ABD [50]	ICCV2021	95.11±0.3	74.83±0.5	46.76±0.6	72.23
CLS-ER [3]	ICLR2022	93.53±0.3	72.11±0.5	57.36±0.7	74.33
FAS [36]	ICLR2022	90.89±1.3	70.89±0.6	60.10±0.2	73.96
DCPOC [53]	PR2023	90.43±0.3	72.20±0.2	53.08±0.2	71.90
PRAKA [47]	ICCV2023	83.74±0.5	76.21±0.4	63.50±0.3	74.48
MIND [4]	AAAI2024	**95.67±0.7**	77.33±0.5	63.82±0.5	78.94
RP^{2}F (ours)	-	91.65±0.3	**83.06±0.2**	**65.81±0.6**	**80.17**

Table 2: Ablation experiment results (ACC %) of RP^{2}F on CIFAR-100 and Tiny-ImageNet.

Hessian estimation	robustness priori	10-split CIFAR-100	10-split Tiny-ImageNet
Identity matrix	×	79.42±0.66	63.96±0.30
Identity matrix	✓	79.91±0.28	64.62±0.38
Fisher information matrix	×	82.29±0.16	64.40±0.57
Fisher information matrix	✓	82.85±0.36	64.79±0.47
Parameter-perturbation based (ours)	×	82.37±0.49	64.78±0.70
Parameter-perturbation based (ours)	✓	**83.06±0.2**	**65.81±0.6**

5.3 Ablation Study

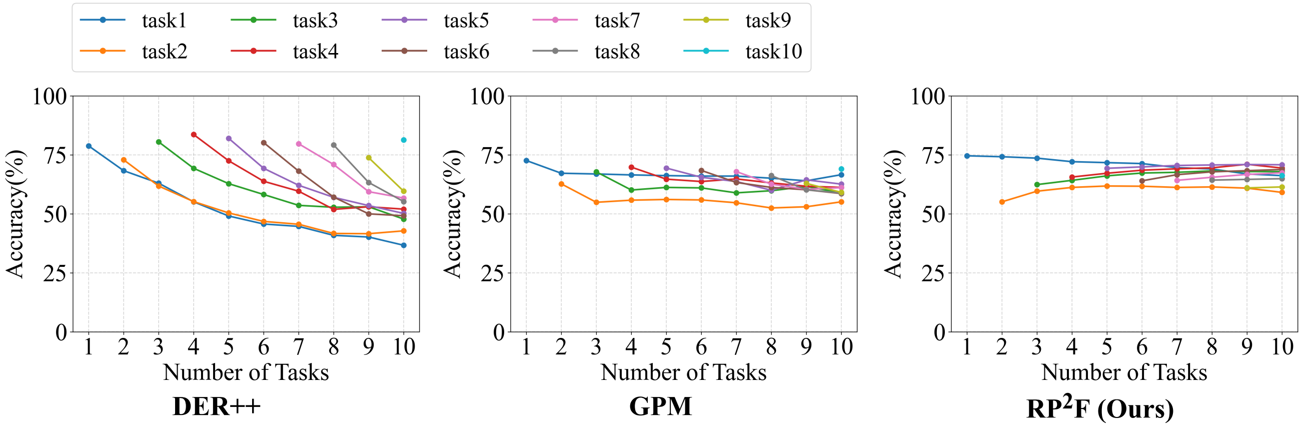
This section delves into the ablation study to analyse the effectiveness of the parameter-perturbation-based Hessian estimation method and the parameter-robustness priori. Specifically, we explore three Hessian estimation strategies: the identity matrix, the Fisher information matrix, and the proposed parameter-perturbation-based approach. For each strategy, we further investigate the impact of introducing the parameter-robustness priori. The results on CIFAR-100 and Tiny-ImageNet are presented in Table 2.

We begin with the identity matrix as the simplest form of Hessian approximation, which yields an accuracy of 79.42% on CIFAR-100 and 63.96% on Tiny-ImageNet. Equipping the parameter-robustness priori provides a slight enhancement, improving accuracies to 79.91% and 64.62%, respectively. Shifting to the widely used Fisher information matrix, we observe a notable improvement in performance. Without the regularizer, it achieves 82.29% accuracy on CIFAR-100 and 64.40% on Tiny-ImageNet. Integrating the parameter-robustness priori further boosts the performance to 82.85% and 64.79%, respectively. Our novel parameter-perturbation-based Hessian estimation method is more competitive than the Fisher information matrix. Without the regularizer, it attains an accuracy of 82.37% on CIFAR-100 and 64.78% on Tiny-ImageNet. The incorporation of the parameter-robustness priori pushes the performance to the highest accuracy: 83.06% on CIFAR-100 and 65.81% on Tiny-ImageNet.

All the results underscore the efficacy of parameter-perturbation-based Hessian estimation in accurately reflecting the true landscape of the loss function. Furthermore, these findings emphasize the ability of the parameter-robustness priori to enhance the model's robustness and mitigate forgetting.

5.4 Analysis of Backward Knowledge Transfer

This section analyzes the backward knowledge transfer of RP^{2}F, a critical IL ability to retain or even enhance performance on previously learned tasks when new tasks are introduced. Our discussion is based on empirical evidence in Figure 2 and Table 3.



DER++ GPM RP^{2}F (Ours)

Figure 2: Accuracy trends over tasks on Tiny-ImageNet for DER++ (left), GPM (middle), and our proposed RP^{2}F model (right).

Table 3: Backward knowledge transfer (BWT, %) on CIFAR-100 and Tiny-ImageNet.

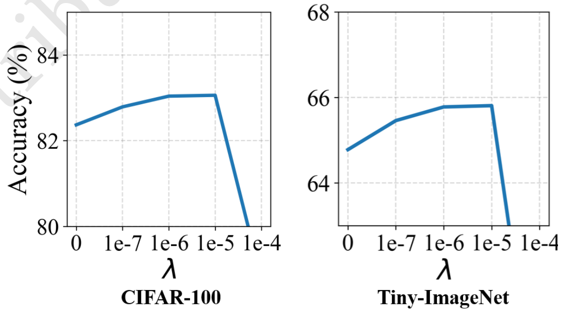
Methods	DER++ [7]	LWF [28]	OWM [67]	PASS [70]	GPM [41]	Adam-NSCL [62]	PRAKA [47]	RP^{2}F (ours)
CIFAR-100	-12.56±5.7	-20.91±3.6	-11.25±0.9	-9.34±0.7	-3.52±0.3	-2.24±0.3	-4.09±0.3	**0.56±0.2**
Tiny-ImageNet	-29.06±1.2	-23.84±1.6	-3.30±0.2	-3.23±0.4	-7.91±0.5	-5.58±0.6	-5.42±0.1	**0.92±0.8**

Figure 2 visually illustrates the performance trajectory of each task through IL. As can be seen, DER++ exhibits a declining trend in accuracy for earlier learned tasks, signifying the occurrence of forgetting. In contrast, GPM maintains a more stable performance throughout the learning phases, showing its effectiveness in mitigating forgetting. Remarkably, RP^{2}F exhibits improved performance on older tasks with the introduction of new tasks (with the exception of the first task).

Further quantitative evidence of BWT is presented in Table 3. The BWT of several methods are calculated on CIFAR-100 and Tiny-ImageNet, including DER++, LwF, OWM, PASS, GPM, Adam-NSCL, and RP^{2}F. It is evident that all baseline methods exhibit a negative BWT, highlighting the challenge of forgetting in IL. In contrast, RP^{2}F records a positive BWT of 0.56% on CIFAR-100 and 0.92% on Tiny-ImageNet, surpassing all other methods. Both visual illustration and quantitative values of BWT underscore RP^{2}F's superior ability to backward knowledge transfer.

5.5 Sensitivity Analysis of λ

This section provides a sensitivity analysis of the hyper-parameter λ, responsible for weighting the parameter-robustness priori in Eq. (9). We conduct experiments with various values of λ and summarize the results in Figure 3. As can be seen, setting λ to zero or an excessively high value results in suboptimal performance of RP^{2}F, attributing to the ineffectiveness or overwhelming impact of the regularizer, which in turn perturbs the training dynamics. Nevertheless, a stable performance window for λ is observed within the range of $[1e-7, 1e-5]$. Based on these empirical findings, we recommend setting λ to $1e-5$ on CIFAR-100 and Tiny-ImageNet to achieve the best overall performance.



CIFAR-100 Tiny-ImageNet

Figure 3: Accuracy of RP^{2}F with various values of λ (Cf. Eq.(9)) on CIFAR-100 (left) and Tiny-ImageNet (right).

6 CONCLUSION

In this work, we propose the Robust Parameter Posterior Fusion (RP^{2}F) framework for incremental learning. RP^{2}F models and fuses the parameter posterior distributions from both new and existing tasks, facilitating a more equitable integration of knowledge across tasks. Furthermore, we incorporate robustness priori into RP^{2}F to mitigate potential forgetting induced by posterior fusion. By employing MAP estimation on the fused posterior, RP^{2}F achieves a harmonious knowledge integration.

RP^{2}F has certain limitations, such as relying on task boundaries to perform posterior fusion. In future work, we plan to address this issue by incorporating online learning techniques to extend the applicability of our method.

REFERENCES

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *European Conference on Computer Vision*. 139–154.

[2] Cem Anil, James Lucas, and Roger Grosse. 2019. Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, Vol. 97. 291–301.

[3] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. 2022. Learning fast, learning slow: a general continual learning method based on complementary learning system. In *International Conference on Learning Representations*. https://openreview.net/forum?id=uxxFrDwrE7Y

[4] Jacopo Bonato, Francesco Pelosin, Luigi Sabetta, and Alessandro Nicolosi. 2024. MIND: Multi-task incremental network distillation. (2024).

[5] Lorenzo Bonicelli, Matteo Boschini, Angelo Porrello, Concetto Spampinato, and Simone Calderara. 2022. On the effectiveness of lipschitz-driven rehearsal in continual learning. In *Advances in Neural Information Processing Systems*.

[6] Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. 2023. Class-incremental continual learning into the eXtended DER-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 5 (2023), 5497–5512.

[7] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and SIMONE CALDERARA. 2020. Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems*, Vol. 33. 15920–15930.

[8] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. 2022. New insights on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations*.

[9] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*.

[10] Hung-Jen Chen, An-Chieh Cheng, Da-Cheng Juan, Wei Wei, and Min Sun. 2020. Mitigating Forgetting in Online Continual Learning via Instance-Aware Parameterization. In *Advances in Neural Information Processing Systems*, Vol. 33. 17466–17477.

[11] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. 2017. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, Vol. 70. 854–863.

[12] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[13] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. 2019. Learning without memorizing. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[14] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. 2020. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision*. 86–102.

[15] Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. 2022. Rank diminishing in deep neural networks. In *Advances in Neural Information Processing Systems*, Vol. 35. 33054–33065.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[17] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211* (2013).

[18] Matthias Hein and Maksym Andriushchenko. 2017. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, Vol. 30.

[19] Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. 2019. Compacting, picking and growing for unforgetting continual learning. In *Advances in Neural Information Processing Systems*, Vol. 32. 13647–13657.

[20] Ta-Chu Kao, Kristopher Jensen, Gido van de Ven, Alberto Bernacchia, and Guillaume Hennequin. 2021. Natural continual learning: success is a journey, not (just) a destination. In *Advances in Neural Information Processing Systems*, Vol. 34. 28067–28079.

[21] Sanghwan Kim, Lorenzo Noci, Antonio Orvieto, and Thomas Hofmann. 2023. Achieving a better stability-plasticity trade-Off via auxiliary networks in continual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 11930–11939.

[22] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* 114, 13 (2017), 3521–3526.

[23] Tatsuya Konishi, Mori Kurokawa, Chihiro Ono, Zixuan Ke, Gyuhak Kim, and Bing Liu. 2023. Parameter-level soft-masking for continual learning. In *International Conference on Machine Learning*, Vol. 202. 17492–17505.

[24] Alex Krizhevsky. 2012. Learning multiple layers of features from tiny images. *University of Toronto* (05 2012).

[25] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Overcoming catastrophic forgetting by incremental moment matching. In *Advances in Neural Information Processing Systems*, Vol. 30.

[26] Xiaorong Li, Shipeng Wang, Jian Sun, and Zongben Xu. 2023. Memory efficient data-free distillation for continual learning. *Pattern Recognition* 144 (2023), 109875.

[27] Xiaorong Li, Shipeng Wang, Jian Sun, and Zongben Xu. 2023. Variational data-free knowledge distillation for continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (2023), 1–17.

[28] Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 12 (2017), 2935–2947.

[29] Yu Liu, Sarah Parisot, Gregory Slabaugh, Xu Jia, Ales Leonardis, and Tinne Tuytelaars. 2020. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. In *European Conference on Computer Vision*. 699–716.

[30] David Lopez-Paz and Marc' Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, Vol. 30. 6467–6476.

[31] David J. C. MacKay. 1992. A practical bayesian framework for backpropagation networks. *Neural Computation* 4, 3 (05 1992), 448–472.

[32] Simone Magistri, Tomaso Trinci, Albin Soutif, Joost van de Weijer, and Andrew D. Bagdanov. 2024. Elastic feature consolidation for cold start exemplar-free incremental learning. In *International Conference on Learning Representations*.

[33] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. 2022. Online continual learning in image classification: An empirical survey. *Neurocomputing* 469 (2022), 28–51.

[34] Arun Mallya and Svetlana Lazebnik. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7765–7773.

[35] Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*. Vol. 24. Elsevier, 109–165.

[36] Zichen Miao, Ze Wang, Wei Chen, and Qiang Qiu. 2022. Continual learning with filter atom swapping. In *International Conference on Learning Representations*.

[37] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. 2018. Variational continual learning. In *International Conference on Learning Representations*.

[38] Quang Pham, Chenghao Liu, and Steven Hoi. 2021. DualNet: continual learning, fast and slow. In *Advances in Neural Information Processing Systems*, Vol. 34. 16131–16144.

[39] Hippolyt Ritter, Aleksandar Botev, and David Barber. 2018. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, Vol. 31.

[40] Grzegorz Rypeść, Sebastian Cygert, Valeriya Khan, Tomasz Trzcinski, Bartosz Michał Zieliński, and Bartłomiej Twardowski. 2024. Divide and not forget: Ensemble of selectively trained experts in continual learning. In *International Conference on Learning Representations*.

[41] Gobinda Saha, Isha Garg, and Kaushik Roy. 2021. Gradient projection memory for continual learning. In *International Conference on Learning Representations*.

[42] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. 2023. Sparse coding in a dual memory system for lifelong learning. In *AAAI Conference on Artificial Intelligence*, Vol. 37. 9714–9722.

[43] Dayana Savostianova, Emanuele Zangrando, Gianluca Ceruti, and Francesco Tudisco. 2023. Robust low-rank training via approximate orthonormal constraints. In *Advances in Neural Information Processing Systems*, Vol. 36. 66064–66083.

[44] Tom Schaul, Sixin Zhang, and Yann LeCun. 2013. No more pesky learning rates. In *International Conference on Machine Learning*, Vol. 28. 343–351.

[45] Tom Schaul, Sixin Zhang, and Yann LeCun. 2013. No more pesky learning rates. In *International Conference on Machine Learning*, Vol. 28. 343–351.

[46] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*. 4528–4537.

[47] Wuxuan Shi and Mang Ye. 2023. Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning. In *IEEE International Conference on Computer Vision*. 1772–1781.

[48] Yujun Shi, Kuangqi Zhou, Jian Liang, Zihang Jiang, Jiashi Feng, Philip H.S. Torr, Song Bai, and Vincent Y. F. Tan. 2022. Mimicking the oracle: An initial phase decorrelation approach for class incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 16722–16731.

[49] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, Vol. 30.

[50] James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. 2021. Always be dreaming: A new approach for data-free class-incremental learning. In *IEEE International Conference on Computer Vision*.

[51] Stanford. 2015. Tiny imageNet challenge (CS231n). http://tiny-imagenet.herokuapp.com/. (2015). http://tiny-imagenet.herokuapp.com/

[52] Wenju Sun, Qingyong Li, Wen Wang, and Yangli-ao Geng. 2023. Towards plastic and stable exemplar-free incremental learning: A dual-learner framework with cumulative parameter averaging. *arXiv preprint arXiv:2310.18639* (2023).

[53] Wenju Sun, Qingyong Li, Jing Zhang, Danyu Wang, Wen Wang, and YangLi ao Geng. 2023. Exemplar-free class incremental learning via discriminative and comparable parallel one-class classifiers. *Pattern Recognition* 140 (2023), 109561.

[54] Wenju Sun, Qingyong Li, Jing Zhang, Wen Wang, and Yangli-ao Geng. 2023. Decoupling learning and remembering: A bilevel memory rramework with knowledge projection for task-incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[55] Wenju Sun, Jing Zhang, Danyu Wang, Yangli-ao Geng, and Qingyong Li. 2021. ILCOC: An incremental learning framework based on contrastive one-class classifiers. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 3580–3588.

[56] Filip Szatkowski, Mateusz Pyla, Marcin Przewięźlikowski, Sebastian Cygert, Bartłomiej Twardowski, and Tomasz Trzciński. 2024. Adapt your teacher: Improving knowledge distillation for exemplar-free continual learning. In *IEEE Winter Conference on Applications of Computer Vision*. 1977–1987.

[57] Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. 2020. Topology-preserving class-incremental learning. In *European Conference on Computer Vision*. Springer, 254–270.

[58] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. 2018. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, Vol. 31.

[59] Gido M Van de Ven and Andreas S Tolias. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734* (2019).

[60] Vinay Kumar Verma, Kevin J Liang, Nikhil Mehta, Piyush Rai, and Lawrence Carin. 2021. Efficient feature transformations for discriminative and generative continual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 13865–13875.

[61] Liyuan Wang, Mingtian Zhang, Zhongfan Jia, Qian Li, Chenglong Bao, Kaisheng Ma, Jun Zhu, and Yi Zhong. 2021. AFEC: Active forgetting of negative transfer in continual learning. In *Advances in Neural Information Processing Systems*, Vol. 34. 22379–22391.

[62] Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. 2021. Training networks in null space of feature covariance for continual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 184–193.

[63] Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Tiehang Duan, and Mingchen Gao. 2022. Improving task-free continual learning by distributionally robust memory evolution. In *International Conference on Machine Learning*, Vol. 162. 22985–22998.

[64] Ju Xu and Zhanxing Zhu. 2018. Reinforced continual learning. In *Advances in Neural Information Processing Systems*, Vol. 31. 907–916.

[65] Hongxu Yin, Pavlo Molchanov, Jose M. Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. 2020. Dreaming to distill: data-free knowledge transfer via deepinversion. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[66] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. 2018. Lifelong learning with dynamically expandable networks.. In *International Conference on Learning Representations*.

[67] Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. 2019. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence* 1, 8 (2019), 364–372.

[68] Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*. 3987–3995.

[69] Fei Zhu, Zhen Cheng, Xu-yao Zhang, and Cheng-lin Liu. 2021. Class-incremental learning via dual augmentation. In *Advances in Neural Information Processing Systems*, Vol. 34. 14306–14318.

[70] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. 2021. Prototype augmentation and self-supervision for incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5871–5880.

[71] Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. 2022. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9296–9305.