
Appendix for Dynamic Siamese Expansion Framework for Improving Robustness in Online Continual Learning

Contents

A	Additional information for related works	2
B	Algorithm pseudocode	3
C	Additional information for the experiment setting	4
C.1	Training details of DSEF	4
C.2	Datasets	4
C.3	Baselines	5
D	Additional ablation studies	6
D.1	Analysis of computational efficiency	6
D.2	Long-Term Stability and Static Branch Decay Analysis	6
D.3	Stability of Mutual Information Estimates (α_k)	6
E	Broader Impact	7

A Additional information for related works

Online Continual Learning (OCL) is an evolving paradigm designed to address the pressing demands of real-time learning systems where data is received sequentially and the underlying distributions shift over time. Unlike traditional batch learning that assumes availability of the entire dataset upfront, OCL algorithms are forced to adapt continually with restricted memory and computational resources [2]. This constraint necessitates a delicate balance between two competing objectives: acquiring new knowledge efficiently and retaining previously learned information to prevent catastrophic forgetting. Early efforts in OCL primarily focused on rehearsal-based techniques, where a small buffer of past data was maintained to refresh the model’s memory, but these approaches struggled with scalability and domain shifts [16]. More recently, architectural and optimization-based mechanisms have been developed to address these challenges more flexibly. Architectural methods often incorporate expandable network modules or dynamic parameter allocation strategies, allowing models to grow or specialize in response to new tasks without overwriting critical representations [20, 4]. Optimization-driven methods include meta-learning approaches and adaptive regularization techniques that modulate plasticity dynamically based on task similarity or uncertainty [13, 12]. Moreover, practical variants of OCL attempt to mirror real-world scenarios more faithfully by decoupling the arrival of inputs and their corresponding labels—such as in robotics or healthcare—where labels may be delayed, sparse, or noisy [13]. Benchmarking OCL methods has also gained attention, with recent initiatives designing evaluation protocols that incorporate realistic distributional shifts, including abrupt changes and non-stationary sequences, which better simulate operational environments [12]. However, a major deficiency in most current OCL frameworks lies in their insufficient consideration of robustness. Real-world deployments, especially in safety-critical domains like autonomous driving and cybersecurity, expose models to adversarial manipulations and unforeseen environmental variations, which many OCL methods fail to robustly address [19]. The gap between continual adaptability and robust performance remains a central challenge and an active area for future research.

The issue of adversarial robustness has dominated the machine learning security discourse for years, evolving from heuristic, often ad-hoc solutions towards more principled and theoretically grounded defenses. Initial attempts at mitigating adversarial attacks included input preprocessing methods—such as JPEG compression and randomization techniques—as well as ensemble-based approaches that aggregated predictions from multiple independently trained models to dilute the effect of perturbations [21, 1]. While these methods provided a degree of empirical defense, they often faltered against adaptive attacks that specifically target their weaknesses [11]. Adversarial training emerged as the most effective and widely adopted defense mechanism, wherein models are exposed to adversarially perturbed inputs during training to harden their decision boundaries [8, 10]. This approach, despite its computational intensity, has consistently improved robustness against a variety of white-box and black-box attack models. Complementary strategies such as defensive distillation—which smooths model gradients—and robust knowledge distillation—which transfers robustness properties from larger models to smaller ones—have further enriched the defense toolkit [7, 24]. Integrating adversarial robustness within the continual learning framework introduces unique complexities. Continual learning demands model plasticity to absorb new tasks while resisting forgetting, but adversarial robustness requires stability under malicious perturbations, which can resemble structured shifts in data distributions. Recent works have proposed reinterpreting adversarial perturbations as task-like domain shifts, leveraging this insight to embed adversarial training procedures into expanding architectures that maintain robustness across tasks [23]. Feature-level and output-level distillation techniques are employed to transfer and preserve robust representations without compromising the model’s ability to learn incrementally. This emerging line of work is

crucial for constructing lifelong learning systems capable of withstanding adversarial threats in dynamic environments.

Dynamic expansion models represent a prominent direction in lifelong learning, offering structural plasticity by progressively increasing model capacity in response to incoming tasks. Such approaches mitigate catastrophic forgetting by isolating task-specific components—neurons, layers, or subnetworks—thus reducing interference between tasks [4, 15, 18]. This strategy contrasts with fixed-capacity models that rely heavily on regularization or rehearsal to maintain performance. Historically, convolutional neural networks (CNNs) have dominated this space, given their proven success in computer vision tasks. However, the rise of Vision Transformers (ViTs) has ushered in a new era of flexible and scalable architectures for continual learning. ViTs leverage self-attention mechanisms to capture global dependencies and can be naturally extended with modular components tailored for incremental learning [5, 6]. Modern ViT-based methods incorporate task-specific attention modules or decoupled task heads to preserve performance on previously learned tasks while facilitating smooth integration of new knowledge [6, 22, 14]. Beyond pure vision architectures, hybrid models that integrate ViTs with large multimodal language models are gaining traction, aiming to enhance both task transferability and domain generalization in lifelong learning settings [17]. Such models combine rich visual and semantic information streams, offering a more comprehensive foundation for adaptation. Despite these architectural advancements, the vast majority of dynamic expansion techniques prioritize mitigating forgetting and enhancing plasticity, often neglecting the equally vital challenge of adversarial robustness and resilience to distributional shifts. Addressing these gaps requires novel frameworks that synergize architectural dynamism with robust learning principles, ensuring lifelong learners are both adaptable and secure in real-world deployments.

Although existing works in online continual adversarial defense have contributed valuable insights, they suffer from several notable limitations that hinder their effectiveness in dynamic, real-world scenarios. Many approaches utilize static or monolithic backbone architectures, which inadequately capture the rich multi-scale representations needed to maintain plasticity and robust adaptation across tasks, especially under adversarial perturbations. Moreover, previous methods often overlook the importance of regulating the optimization process to selectively constrain shifts in both predictions and internal representations of historical experts, resulting in substantial forgetting of previously acquired robust knowledge. Additionally, knowledge integration from prior tasks tends to rely on heuristic or uniform fusion strategies without rigorously measuring the relevance between historical experts and the current task, which may lead to negative transfer or inefficient exploitation of useful prior information. In contrast, our paper advances the field by proposing a Dynamic Siamese Expert Fusion (DSEF) framework that effectively manages a Siamese backbone to capture complementary global and local representations, thereby significantly enhancing the model’s plasticity and robustness. Complementing this, our Robust Dynamic Representation Optimization (RDRO) carefully controls optimization dynamics to selectively minimize prediction and representation shifts for each historical expert, preventing forgetting while preserving previously learned robust abilities under adversarial conditions. Furthermore, the proposed Robust Feature Fusion (RFF) employs a mutual information criterion to quantitatively evaluate knowledge similarity between each history expert and the new task, enabling adaptive, weighted fusion of features that promotes positive knowledge transfer and reduces negative interference. Collectively, these innovations overcome the deficiencies of earlier methods by providing a more flexible, adaptive, and theoretically grounded framework that robustly integrates past knowledge while efficiently learning new tasks under evolving adversarial environments, thus significantly improving lifelong learning performance in online continual adversarial defense settings.

B Algorithm pseudocode

Algorithm 1 The learning process of the proposed DSEF.

Input: The number of tasks (N), a series of tasks $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_N\}$.

Output: The model’s parameter set

```
1: for  $i < N$  do
2:   Step 1: Model expansion process:
3:   if  $i = 0$  then
4:     Build the foundational backbone  $F_{\theta^a}$ 
5:     Build the dynamic and static backbones  $F_{\theta^d}$  and  $F_{\theta^s}$ , respectively.
6:     Build the first expert  $\mathcal{E}_1 = \{F_{\varphi_1^f}, F_{\varphi_1^c}, F_{\gamma_1}\}$ 
7:   else
8:     Build a new expert  $\mathcal{E}_i = \{F_{\varphi_i^f}, F_{\varphi_i^c}, F_{\gamma_i}\}$ 
9:   end if
10:  Step 2: Calculate robust optimization loss terms:
11:  for  $j < n'$  do
12:    Get the data batch  $(\mathcal{X}, \mathcal{Y})$  from  $\mathcal{C}_i$ 
13:    Calculate  $F'_{\text{pre}}$  using Eq.6
14:    Calculate  $F_{\text{feature}}$  using Eq.14
15:    Step 3: Mutual information fusion:
16:    Calculate the mutual information weights using Eq.17
17:    Calculate the augmented representations using Eq.18
18:    Get the final prediction using Eq.19
19:    Step 4: Optimizing the model’s parameters:
20:    Update the model’s parameters using Eq.20
21:  end for
22: end for
```

C Additional information for the experiment setting

C.1 Training details of DSEF

We utilize one Vision Transformer (ViT) as the dynamic backbone and another ViT as the static backbone. To enhance feature extraction, we adopt a Siamese mechanism that effectively combines the representational strengths of both backbones. The network parameters are optimized using the Adam optimizer with a learning rate of 0.03 and a batch size of 32.

C.2 Datasets

To comprehensively assess the proposed DSEF model, we conduct experiments on four well-established benchmark datasets commonly used in image classification tasks. First, CIFAR-10 consists of 60,000 tiny color images (32×32 pixels) evenly split across 5 distinct object categories, providing a standard baseline for classification performance. The more granular CIFAR-100 dataset, also containing 60,000 images of the same resolution, extends this challenge by offering 100 fine-grained classes organized into 10 broader superclasses, enabling evaluations at multiple semantic levels. We also include the CUB-200 dataset, a specialized fine-grained collection focusing on bird species, which contains nearly 12,000 images across 200 classes, offering a domain with subtle inter-class variations. Lastly, TinyImageNet is employed as a computationally efficient subset derived from the larger ImageNet corpus, comprising 200 categories with a balanced number of training, validation, and test images, all resized to 64×64 pixels. This diverse selection of datasets ensures that the evaluation covers a range of classification difficulties, from general object recognition to fine-grained categorization, and varying dataset scales.

Methods	Train Params↓	Train Time↓
Refresh	86.30M	09.20m
Refresh(Adv)	86.30M	16.51m
DER	86.30M	09.20m
DER(Adv)	86.30M	16.51m
DER++	86.30M	09.20m
DER++(Adv)	86.30M	16.51m
AIR	86.30M	04.91m
DSEM	91.26M	07.63m

Table 1: Comparison of our method with other SOTA methods in terms of training parameters and training time under the TinyImageNet dataset. Arrows (\uparrow , \downarrow) indicate whether higher or lower values are preferred for each metric.

C.3 Baselines

In this section, we provide a comprehensive comparison between our proposed method and several well-established continual learning baselines, with a particular emphasis on experience replay-based techniques, which have demonstrated strong performance in mitigating catastrophic forgetting by storing and revisiting past data. The first baseline, Refresh [9], employs a straightforward replay mechanism combined with regularization to maintain stability of learned representations. While Refresh is effective in preserving prior knowledge through replay buffers, its reliance on a fixed backbone network limits its adaptability to new tasks or distributional shifts, especially in adversarial contexts. Building upon this, DER and DER++ [3] introduce distillation-enhanced replay strategies, where knowledge distillation is used alongside replay to better align feature representations over time and reduce forgetting. DER++ further improves on DER by incorporating more refined distillation losses and exemplar selection methods, enhancing stability and plasticity trade-offs. However, both DER and DER++ still operate with static backbones, which restrict their flexibility in accommodating task-specific variations and adversarial perturbations.

Given that our framework integrates adversarial training to bolster robustness, it is crucial to evaluate how these baselines perform under adversarial conditions. Thus, we also consider their adversarially trained variants: Refresh (Adv), DER (Adv), and DER++ (Adv). These variants apply adversarial perturbations during training to improve model resilience against attacks. Although adversarial training improves robustness to some extent, these methods often suffer from increased computational overhead and may still experience degraded performance when facing evolving or stronger adversarial threats in continual learning settings. The static nature of their backbone architectures further exacerbates the challenge of balancing plasticity and robustness over time.

Additionally, we incorporate AIR [23], a novel method specifically developed for continual adversarial defense. AIR addresses the challenge of adapting to a sequence of evolving adversarial attacks by introducing an innovative replay mechanism that balances model plasticity and stability. It employs isotropic replay to maintain consistency in the local data distribution, aligning the model’s predictions between past and new tasks, while anisotropic replay allows the model to capture a richer, mixed semantic representation to better prepare for future attacks. Complementing these, AIR integrates a simple yet effective regularizer that further alleviates the trade-off between retaining previous defenses and learning new ones. Experimental results show that AIR can achieve performance comparable to, or even surpassing, the ideal scenario of joint training with access to all attack data, making it a strong baseline for lifelong adversarial defense.

Overall, these baselines represent a diverse spectrum of continual learning strategies, ranging from simple replay to advanced distillation and dynamic expansion mechanisms, each with their respective strengths and limitations. Our comparative analysis highlights the need for approaches that not

Task Index	Static Only (%)	Dynamic Only (%)	Full Model (%)	Static vs. Dynamic Gap (%)
Task 1	67.6	68.0	68.17	-0.4
Task 5	65.7	67.1	68.05	-1.4
Task 10	63.6	66.7	67.95	-3.1
Task 15	61.1	65.6	67.65	-4.5
Task 20	60.1	65.9	67.38	-5.8

Table 2: Long-term stability analysis on Split CIFAR-100 (20 tasks).

only prevent forgetting but also maintain robustness and adaptability in adversarial, non-stationary environments—a gap our proposed method aims to fill effectively.

D Additional ablation studies

D.1 Analysis of computational efficiency

We conduct a comparative analysis of the computational cost of our approach and several baselines on the Split TinyImageNet dataset. The evaluation considers both the number of trainable parameters (in millions) and the total training time (in minutes), as summarized in Tab. 1. Results show that our method introduces only a marginal increase in model size while maintaining a practical training cost. In contrast, static baselines such as DER(Adv) and DER++(Adv) exhibit significantly higher computational time during training.

D.2 Long-Term Stability and Static Branch Decay Analysis

To further assess the complementary roles and stability of different components in our framework, we conduct an additional analysis on a 20-task Split CIFAR-100 sequence.

We compare three variants: (1) **Static Only**: model using only the semi-static anchor branch; (2) **Dynamic Only**: model using only the ensemble of dynamic experts; (3) **Full Model**: our proposed model combining both static and dynamic components.

As shown in Table 2, the static branch exhibits a gradual accuracy decay from 67.6% (Task 1) to 60.1% (Task 20), corresponding to a nonlinear but bounded degradation (5.8%) over 20 tasks. This degradation is compensated by the dynamic experts, leading to stable performance in the full model (68.17 \rightarrow 67.38%). Notably, the fusion mechanism (KRA) mitigates potential loss of expressivity by re-weighting task-relevant experts, maintaining both long-term stability and adaptability.

D.3 Stability of Mutual Information Estimates (α_k)

To further understand the reliability of the fusion mechanism, we conduct a statistical analysis of the mutual information weights α_k across tasks on CIFAR-100. We evaluate four metrics: the maximum value $\max \alpha_k$, standard deviation $\text{Std}(\alpha_k)$, entropy H , and the number of dominant experts activated per task. As shown in Table 3, α_k tends to concentrate on a few experts while maintaining a smooth distribution across tasks.

Observations. We find that α_k distributions are stable and interpretable: (1) α_k typically concentrates on 1-3 experts, avoiding over-fragmentation of task routing; (2) incorporating softmax with label smoothing mitigates mutual information overestimation under class imbalance; and (3) entropy trends reflect a meaningful knowledge transfer distribution, rather than degenerate or collapsed α_k assignments.

Table 3: Statistical analysis of mutual information weights

Task ID	Max α_k	Std (α_k)	Entropy (H)	Dominant Expert Count
T1	1.00	0.00	0.00	1
T3	0.54	0.12	0.98	2
T6	0.43	0.17	1.14	3
T9	0.36	0.22	1.27	3-4

E Broader Impact

This work presents a new framework for continual learning that effectively mitigates catastrophic forgetting while strengthening resilience against adversarial attacks. The proposed approach is well-suited for deployment in real-world scenarios where machine learning models must maintain reliable performance over time, even under evolving and potentially hostile conditions. Examples include autonomous driving systems, healthcare monitoring technologies, and security-critical infrastructures.

By emphasizing robustness and long-term reliability, our method contributes to the development of safer and more trustworthy machine learning applications in sensitive domains. In addition, we provide a comprehensive benchmark and evaluation strategy that can support the broader research community in building more resilient continual learning algorithms.

At the same time, we acknowledge that such powerful capabilities come with important responsibilities. Without proper governance, continual learning techniques could be misapplied in areas such as surveillance or defense. Furthermore, although the proposed methods improve robustness, they do not guarantee immunity to all types of attacks. To mitigate these concerns, we advocate for careful validation in safety-critical settings and strict adherence to ethical standards throughout the development and deployment process.

References

- [1] Alexander Bagnall, Razvan Bunescu, and Gordon Stewart. Training ensembles to detect adversarial examples. *arXiv preprint arXiv:1712.04006*, 2017.
- [2] Seyed Amir Bidaki, Amir Mohammadkhah, Kiyan Rezaee, Faeze Hassani, Sadegh Eskandari, Maziar Salahi, and Mohammad M. Ghassemi. Online continual learning: A systematic literature review of approaches, challenges, and benchmarks, 2025.
- [3] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.
- [4] C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, and S. Yang. Adanet: Adaptive structural learning of artificial neural networks. In *Proc. of Int. Conf. on Machine Learning (ICML)*, vol. PMLR 70, pages 874–883, 2017.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [6] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9285–9295, 2022.

- [7] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3996–4003, 2020.
- [8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [9] Tyler L Hayes, Nathan D Cahill, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *European Conference on Computer Vision (ECCV)*, pages 466–483. Springer, 2020.
- [10] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Las-at: adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13398–13408, 2022.
- [11] Guoqing Jin, Shiwei Shen, Dongming Zhang, Feng Dai, and Yongdong Zhang. Ape-gan: Adversarial perturbation elimination with gan. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3842–3846. IEEE, 2019.
- [12] Dahuin Jung, Dongjin Lee, Sunwon Hong, Hyemi Jang, Ho Bae, and Sungroh Yoon. New insights for the stability-plasticity dilemma in online continual learning. In *International Conference on Learning Representations*, 2023.
- [13] Binhong Liu, Dexin Yao, Rui Yang, Zhi Yan, and Tao Yang. Semi-supervised online continual learning for 3d object detection in mobile robotics. *Journal of Intelligent & Robotic Systems*, 110(4):1–16, 2024.
- [14] Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [15] R. Polikar, L. Upda, S. S. Upda, and Vasant Honavar. Learn++: An incremental learning algorithm for supervised neural networks. *IEEE Trans. on Systems Man and Cybernetics, Part C*, 31(4):497–508, 2001.
- [16] Ameya Prabhu, Zhipeng Cai, Puneet K. Dokania, Philip H. S. Torr, Vladlen Koltun, and Ozan Sener. Online continual learning without the storage constraint. *CoRR*, abs/2305.09253, 2023.
- [17] Biqing Qi, Xinquan Chen, Junqi Gao, Dong Li, Jianxing Liu, Ligang Wu, and Bowen Zhou. Interactive continual learning: Fast and slow thinking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12892, 2024.
- [18] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.
- [19] Albin Soutif-Cormerais, Antonio Carta, Andrea Cossu, Julio Hurtado, Vincenzo Lomonaco, Joost Van de Weijer, and Hamed Hemati. A comprehensive empirical evaluation on online continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3518–3528, October 2023.
- [20] Maorong Wang, Nicolas Michel, Ling Xiao, and Toshihiko Yamasaki. Improving plasticity in online continual learning via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23460–23469, June 2024.

- [21] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.
- [22] Mengqi Xue, Haofei Zhang, Jie Song, and Mingli Song. Meta-attention for vit-backed continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 150–159, 2022.
- [23] Yuhang Zhou and Zhongyun Hua. Defense without forgetting: Continual adversarial defense with anisotropic & isotropic pseudo replay. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24263–24272, 2024.
- [24] Jianing Zhu, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, and Hongxia Yang. Reliable adversarial distillation with unreliable teachers. *arXiv preprint arXiv:2106.04928*, 2021.