
Proceedings Paper:
YE, FEI, Zhao, Yulong, Liu, Qihe et al. (2025) Dynamic Siamese Expansion Framework for Improving Robustness in Online Continual Learning. In: NeurIPS 2025: The Thirty-Ninth Annual Conference on Neural Information Processing Systems. Curran Associates Inc..

Dynamic Siamese Expansion Framework for Improving Robustness in Online Continual Learning

Fei Ye[1], **Yulong Zhao**[1], **Qihe Liu**[1*], **Junlin Chen**[1], **Adrian G. Bors**[2]
Jingling Sun[1], **Rongyao Hu**[1], **Shijie Zhou**[1]
[1]School of Information and Software Engineering,
University of Electronic Science and Technology of China
[2]Department of Computer Science, University of York
{ feiye@uestc.edu.cn, yulongzhao913@outlook.com, qiheliu@uestc.edu.cn,
adrian.bors@york.ac.uk, jlsun@uestc.edu.cn, ryhu@uestc.edu.cn,
sjzhou@uestc.edu.cn }

Abstract

Continual learning requires the model to continually capture novel information without forgetting prior knowledge. Nonetheless, existing studies predominantly address catastrophic forgetting, often neglecting enhancements in model robustness. Consequently, these methodologies fall short in real-time applications, such as autonomous driving, where data samples frequently exhibit noise due to environmental and lighting variations, thereby impairing model efficacy and causing safety issues. In this paper, we address robustness in continual learning systems by introducing an innovative approach, the Dynamic Siamese Expansion Framework (DSEF) that employs a Siamese backbone architecture, comprising static and dynamic components, to facilitate the learning of both global and local representations over time. Specifically, the proposed framework dynamically generates a lightweight expert for each novel task, leveraging the Siamese backbone to enable rapid adaptation. A novel Robust Dynamic Representation Optimization (RDRO) approach is proposed to incrementally update the dynamic backbone by maintaining all previously acquired representations and prediction patterns of historical experts, thereby fostering new task learning without inducing detrimental knowledge transfer. Additionally, we propose a novel Robust Feature Fusion (RFF) approach to incrementally amalgamate robust representations from all historical experts into the expert construction process. A novel mutual information-based technique is employed to derive adaptive weights for feature fusion by assessing the knowledge relevance between historical experts and the new task, thus maximizing positive knowledge transfer effects. A comprehensive experimental evaluation, benchmarking our approach against established baselines, demonstrates that our method achieves state-of-the-art performance even under adversarial attacks. Code is released at https://github.com/seSysdl/DSEF.

1 Introduction

Continual/Lifelong Learning (CL) has emerged as a pivotal subject within the domain of deep learning, significantly contributing to the progression of artificial intelligence systems [22]. In contrast to conventional deep learning methodologies, continual learning introduces a unique training framework wherein the model is exposed to a constrained set of samples, with prior data samples rendered inaccessible. This learning paradigm encounters a critical challenge termed catastrophic forgetting

*corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

[27], which can substantially impede the model's efficacy. This deterioration in performance occurs when the model modifies its parameters to assimilate new tasks.

Recent investigations into the mitigation of catastrophic forgetting in continual learning have delineated several strategic approaches, which are predominantly categorized into three principal domains: dynamic expansion methodologies [6, 14], which augment the model's capacity through the dynamic incorporation of additional hidden nodes and layers; memory-based approaches [5, 2], which enhance model efficacy by utilizing a judiciously curated set of samples retained within a memory buffer; and regularization techniques [19, 25], which typically integrate an auxiliary regularization term into the primary objective function to safeguard critical network parameters from substantial alterations. Among these methodologies, memory-based approaches exhibit efficacy in mitigating network forgetting when confronted with a constrained number of tasks, yet they frequently exhibit suboptimal performance in more challenging learning scenarios in which the number of tasks grows over time. Conversely, dynamic expansion methodologies are favored for their scalability and adaptability, rendering them apt for a diverse array of continual learning applications.

The majority of existing continual learning research presupposes that data samples are derived from the original data distribution [37]. Nonetheless, in more pragmatic scenarios such as autonomous driving, data samples frequently exhibit noise due to dynamically fluctuating illuminations, weather conditions, and road surfaces. Such noise-laden data samples can impair model performance, potentially leading to car accidents. This paper aims to enhance model robustness in continual learning by investigating a novel learning paradigm termed Online Continual Adversarial Defense (OCAD), wherein new data samples are encountered only once, and the model is expected to perform proficiently on both clean and adversarial samples post-training. OCAD presents three challenges: the adaptability to novel tasks (plasticity), the retention of antecedently acquired knowledge (stability), and the capability to counter adversarial samples (robustness). These challenges are mutual interaction during the training process, leading to significant performance degeneration for models.

To enhance plasticity, this study introduces an innovative Dynamic Siamese Expansion Framework (DSEF) that orchestrates and refines a Siamese backbone architecture to capture the semantically rich information. As a result, the Siamese backbone can help create a lightweight expert to adapt to a new task. The proposed Siamese backbone architecture comprises a static backbone for capturing global representations across all tasks and a dynamic backbone for delivering local representations, both of which are implemented using a pre-trained Vision Transformer (ViT) [8] to facilitate rapid adaptation. Moreover, the static and dynamic backbones predominantly share parameters to augment communication capabilities and diminish model complexity. Additionally, a learnable strategy network is proposed to ascertain and generate adaptive weights that delineate the significance of the static and dynamic backbones, thereby achieving optimal generalization performance.

To ensure robust stability, this paper introduces an innovative Robust Dynamic Representation Optimization (RDRO) methodology, which incrementally refines the dynamic backbone while preserving the static backbone in a fixed state throughout the optimization process. Specifically, the RDRO methodology formulates the static backbone as an auxiliary model that guides the optimization trajectory of the dynamic backbone through two regularization loss terms. The first loss term assesses the divergence between predictions of historical experts constructed from previously and currently acquired dynamic backbones, which ensures that updating the dynamic backbone does not precipitate substantial alterations in the prediction patterns of each historical expert. The subsequent loss term minimizes statistical discrepancies in the representations generated by previously and currently learned dynamic backbones, thereby preserving previously acquired robust representations.

To enhance adversarial robustness, we propose to integrate adversarial loss terms into the proposed RDRO framework to learn robust representations, ensuring optimal performance on both clean and adversarial samples. In addition, an innovative Robust Feature Fusion (RFF) methodology is introduced to amalgamate all previously acquired robust representations from historical experts with the representation extracted by the current expert, thereby facilitating the learning of new tasks. To optimize the positive transfer knowledge effects, the RFF method evaluates the knowledge similarity between each historical expert and the new task using a mutual information criterion, employing these metrics as adaptive weights in the feature fusion process. This strategy effectively reutilizes unactivated parameters and representations to enhance new task learning, resulting in superior generalization performance. A comprehensive series of experiments conducted across diverse datasets empirically demonstrates that the proposed approach achieves state-of-the-art performance.

The principal contributions of this research are delineated as follows : (1) This paper addresses a novel and challenging OCAD by proposing a novel DSEF that manages a Siamese backbone structure to capture global and local representations, enhancing plasticity; (2) This paper proposes a novel RDRO approach to regulate the optimization behaviour of the dynamic backbone by selectively minimizing the prediction and representation shifts of each history expert, which can prevent forgetting and maintain previously learned robust abilities; (3) This paper proposes a novel RFF approach to integrating all previously learned robust representations to promote the new task learning. Specifically, the proposed RFF approach evaluates the knowledge similarity between each history expert and the new task via a mutual information criterion, which provides adaptive weights for the feature fusion process, leading to better positive knowledge transfer effects.

2 Related Work

Adversarial Defense. Adversarial robustness has become a central concern in machine learning security, leading the field from early-stage heuristic defenses such as input preprocessing, generative noise suppression, and ensemble-based stabilization [35, 17, 1], to more principled and theoretically grounded approaches. Although initial methods offered short-term protection, they often lacked generalizability under adaptive attack scenarios. In contrast, adversarial training, which incorporates perturbed samples during model optimization, has demonstrated strong effectiveness and remains one of the most widely adopted defense strategies [11, 16, 23]. Additional techniques, including defensive distillation and robust knowledge transfer, have further enhanced model resilience against subtle and targeted manipulations [10, 33, 40]. Within the domain of continual learning, combining robustness with plasticity presents unique challenges. Recent studies have begun to explore this intersection by interpreting adversarial perturbations as structured task-like shifts, rather than treating them as isolated threats [39]. Building on this perspective, some methods embed adversarial training into architectures that expand over time, while employing feature and output distillation to prevent forgetting and preserve robustness across evolving tasks. This integrated direction offers a promising foundation for developing more secure and adaptable lifelong learning systems.

Dynamic Expansion Model. Lifelong learning has increasingly leveraged dynamic architectural strategies, where models evolve over time by integrating new neurons, layers, or specialized modules to handle incoming tasks. This structural plasticity allows for continual adaptation while minimizing interference with previously acquired knowledge by isolating task-specific components [6, 15, 28, 30, 34, 38, 18, 32]. Although convolutional neural networks (CNN) have traditionally served as the foundation for such approaches, the growing adoption of Vision Transformers (ViTs) reflects a broader shift toward architectures with greater capacity for scalability and flexible representation learning [8, 9]. Modern methods often incorporate modular attention mechanisms and decoupled task heads within ViT-based frameworks to better support incremental learning without performance degradation on earlier tasks [9, 36, 26]. Additionally, recent developments explore hybrid models that jointly optimize visual transformers with large-scale multimodal language models, aiming to improve both task transfer and generalization in dynamic settings [29]. Despite these innovations, many existing solutions remain primarily focused on preventing forgetting, with limited attention paid to adversarial robustness and resilience to distributional changes. More information can be found in **Appendix-A** from Supplementary Material (SM).

3 Methodology

3.1 Problem Statement

In continual learning, it is presumed that a model has access solely to a limited set of training samples for each task, while previous tasks are not accessible. The main goal of the model is to acquire new information without losing previously learned knowledge. Additionally, instead of most existing studies, which focus on a simple continual learning scenario, this paper explores a more intricate and realistic continual learning scenario referred to as Online Continual Adversarial Defense (OCAD), which introduces adversarial attacks aimed at undermining the model's performance. Consistent with the class-incremental framework, a training dataset $\mathcal{C}^s = \{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^{n^s}$ in the OCAD setting is divided into N subsets $\{\mathcal{C}_1^s, \cdots, \mathcal{C}_N^s\}$ according to the category information, where n^s denotes the total number of training data samples and each subset $\mathcal{C}_i^s$ contains data samples from one or several adjacent classes. $\mathbf{x}_j \in \mathcal{X}$ represents the j-th data sample, and $\mathbf{y}_j \in \mathcal{Y}$ denotes the corresponding class label. $\mathcal{X}$ and $\mathcal{Y}$ signify the data and label spaces, respectively. Each subset $\mathcal{C}_i^s$ is treated as a

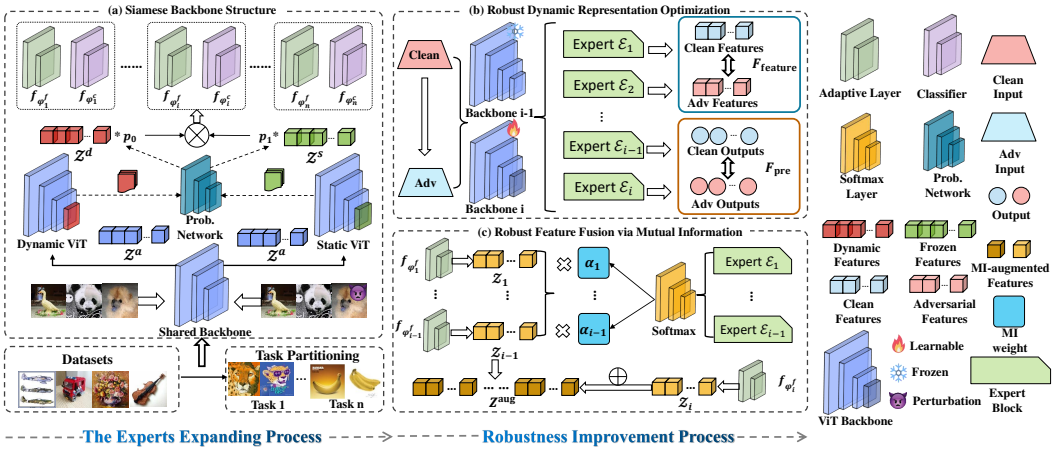


Figure 1: The comprehensive framework of the proposed DSEF including the RDRO and MBRFF mechanisms. The RDRO mechanism aligns the feature distributions and outputs of clean and adversarial samples in the siamese ViT network. Meanwhile, the MBRFF mechanism uses historical experts to augment feature extraction by mutual information.

specific task, denoted as $\mathcal{T}_i$. During the training process of a certain task ($\mathcal{T}_i$), the learning goal of a model is to find an optimal parameter set that minimizes the loss values of all previous and new data samples, expressed as :

$$\theta^\star = \underset{\theta\in\Theta}{\operatorname{argmin}}\Big\{\sum\nolimits_{c=1}^{j}\{\sum\nolimits_{t=1}^{|\mathcal{C}_c^s|}\{F_{\mathrm{ce}}(\mathbf{y}_t, f_\theta(\mathbf{x}_t))\}\Big\}, \tag{1}$$

where Θ denotes the model's parameter space, and $|\mathcal{C}_c^s|$ represents the cardinality of the sample set $\mathcal{C}_c^s$. The cross-entropy loss is computed via the function $F_{\mathrm{ce}}(\cdot)$. The intractability of identifying the optimal solution, as defined by Eq. (1), arises from the unavailability of data samples from all prior tasks. To mitigate this, existing studies have proposed the utilization of a fixed-size memory buffer [5] to preserve and replay critical past examples during the learning phase of a new task. When the new task learning is finished, the model's performance is evaluated across all testing datasets $\{\mathcal{C}_1^T, \cdots, \mathcal{C}_N^T\}$ using classification accuracy as the metric. In the OCAD framework's testing phase, the model's robustness is assessed via various adversarial attack methods, denoted as $\mathcal{A} = \{\mathcal{A}_1, \cdots, \mathcal{A}_T\}$. Each adversarial method $\mathcal{A}_j$ generates an adversarial dataset $\tilde{\mathcal{C}}_{i,j}^T = \mathcal{A}_j(\mathcal{C}_i^T, f_\theta)$ based on a testing dataset $\mathcal{C}_i^T$. The model's robustness is subsequently evaluated across all adversarial datasets $\{\tilde{\mathcal{C}}_{1,1}^T, \cdots, \tilde{\mathcal{C}}_{N,T}^T\}$ using classification accuracy metrics.

3.2 Siamese Backbone Structure

Existing studies in continual learning have investigated the efficacy of leveraging a pre-trained Vision Transformer (ViT) [8] backbone to enhance model performance. These methodologies typically employ a dynamic expansion framework, which utilizes a frozen ViT backbone to facilitate the construction of expert models. This design paradigm effectively preserves prior task knowledge by freezing all parameters of the pre-trained ViT but exhibits limitations in the context of novel task acquisition. To mitigate these constraints, this paper introduces a novel Siamese backbone architecture, which strategically employs both a static and a dynamic backbone to capture static and dynamic information, each instantiated with a pre-trained ViT. The static backbone maintains frozen parameters to furnish a generalized representation applicable across all tasks, whereas the dynamic backbone dynamically optimizes its parameters to adaptive representations. Given the substantial number of hidden layers and parameters inherent in the ViT-based backbone, updating all parameters would incur significant computational overhead. To address this, we propose training only the final K representation layers of the dynamic backbone. Furthermore, the proposed Siamese backbone structure facilitates parameter sharing between the static and dynamic backbones, thereby minimizing redundant parameters and fostering inter-backbone communication.

Let $F_{\theta^a}: \mathcal{X} \to \mathcal{Z}^a$ represent a shared backbone, which processes a data sample $\mathbf{x}$ from the input space $\mathcal{X}^a$ and yields a feature representation $\mathbf{z}^a$ within the feature space $\mathcal{Z}^a$. Furthermore, let $F_{\theta^s}: \mathcal{Z}^a \to \mathcal{Z}$ and $F_{\theta^d}: \mathcal{Z}^d \to \mathcal{Z}$ denote the static and dynamic backbones, respectively, each of

which receives a feature vector extracted by F_{θ^a} and produces a representation $\mathbf{z}$ in the feature space $\mathcal{Z}$. By using the Siamese backbone architecture, an augmented representation can be formulated by :

$$\hat{\mathbf{z}} = F_{\theta^d}(F_{\theta^a}(\mathbf{x})) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x}))\,, \tag{2}$$

where $\otimes$ signifies the concatenation operation, which merges two feature representations. According to Eq. (2), the proposed methodology adaptively generates a novel expert to learn a new task, comprising a fully connected layer $F_{\varphi_j^f}: \mathcal{Z}^2 \to \mathcal{Z}'$ and a linear classifier $F_{\varphi_j^c}: \mathcal{Z}' \to \mathcal{Y}$, where φ_j^f and φ_j^c represent the parameters of the j-th expert. $\mathcal{Z}^2$ denotes the space of $\hat{\mathbf{z}}$ derived via the static and dynamic backbones. Furthermore, $\mathcal{Y}$ signifies the prediction space. The predictive function of the j-th expert is formulated as follows :

$$\mathcal{F}_{\mathrm{p}}(\mathbf{x}, \mathcal{E}_j) = F_{\varphi_j^c}\Big(F_{\varphi_j^f}\Big(F_{\theta^d}(F_{\theta^a}(\mathbf{x})) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x}))\Big)\Big)\,, \tag{3}$$

where $\{y'_1, \cdots, y'_M\} = \mathcal{F}_{\mathrm{p}}(\mathbf{x}, \mathcal{E}_j)$ denotes the predicted probability vector and M is the total number of classes.

Learnable Strategy Network. The representations delineated in Eq. (2) treat the static and dynamic backbones equivalently within the prediction process and thus would not yield optimal performance. Given that the static and dynamic backbones capture global and local representations, it is imperative to ascertain the significance of each representation autonomously, contingent upon the data's inherent characteristics. To this end, this study introduces a novel, learnable strategy network $F_{\gamma_j}: \mathcal{Z}^2 \to \mathcal{Y}'$ with the parameter set γ_j for the j-th expert. This network processes a concatenated feature vector $\hat{\mathbf{z}}$, derived via Eq. (2), and subsequently outputs a selector probability vector over the space $\mathcal{Y}'$. Specifically, the predictive process of the j-th expert, facilitated by the learnable strategy network F_{γ_j}, is formalized as :

$$\mathcal{F}'_{\mathrm{p}}(\mathbf{x}, \mathcal{E}_j) = F_{\varphi_j^c}\Big(F_{\varphi_j^f}\Big(F_{\gamma_j}(\mathbf{x})[0] F_{\theta^d}(F_{\theta^a}(\mathbf{x})) \otimes F_{\gamma_j}(\mathbf{x})[1] F_{\theta^s}(F_{\theta^a}(\mathbf{x}))\Big)\Big)\,, \tag{4}$$

where $F_{\gamma_j}(\mathbf{x})[0]$ and $F_{\gamma_j}(\mathbf{x})[1]$ denote the first and second dimensions of $F_{\gamma_j}(\mathbf{x})$. Compared to Eq. (3), the network F_{γ_j} used in Eq. (4) can yield data-driven adaptive weights that determine the importance of static and dynamic backbones during the prediction process, which can achieve optimal performance.

3.3 Robust Dynamic Representation Optimization

Updating the parameters of the dynamic backbone F_{θ^d} is susceptible to detrimental knowledge transfer effects, given that all historical experts maintain parameter immutability throughout the learning phase of a novel task. To mitigate this, we introduce a novel methodology, termed Robust Dynamic Representation Optimization (RDRO), designed to optimize the dynamic backbone while minimizing catastrophic forgetting. Specifically, the objective of updating the dynamic backbone F_{θ^d} is to promote the new task learning and ensure the preservation of predictive capabilities and robust abilities acquired by each historical expert. To achieve this, the proposed RDRO approach minimizes the divergence between predictions generated using previously and currently acquired representations during the j-th task's learning phase, formally expressed as :

$$\begin{aligned} F_{\mathrm{pre}} = \sum\nolimits_{i=1}^{j-1} \Big\{ & F_{\mathrm{mse}}\Big(F_{\varphi_i^c}\Big(F_{\varphi_i^f}\Big(F_{\theta^d}(F_{\theta^a}(\mathbf{x})) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x}))\Big)\Big), \\ & F_{\varphi_i^c}\Big(F_{\varphi_i^f}\Big(F_{\theta^s}(F_{\theta^a}(\mathbf{x})) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x}))\Big)\Big)\Big\}\,, \end{aligned} \tag{5}$$

where $F_{\mathrm{mse}}(\cdot)$ signifies the Mean Squared Error (MSE) criterion. Nevertheless, the loss function delineated in Eq. (5) exclusively accounts for clean data samples, thereby disregarding adversarial data. Consequently, the dynamic backbone is rendered incapable of preserving the robust representation information. To mitigate this, the proposed RDRO methodology incorporates adversarial loss into Eq. (5), yielding :

$$\begin{aligned} F'_{\mathrm{pre}} = \min_{\theta^d} \Big\{ & F_{\mathrm{pre}}(\mathbf{x}) + \sum\nolimits_{i=1}^{j-1} \Big\{ F_{\mathrm{mse}}\Big(F_{\varphi_i^c}\Big(F_{\varphi_i^f}\Big(F_{\theta^d}(F_{\theta^a}(\mathbf{x})) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x}))\Big)\Big), \\ & F_{\varphi_i^c}\Big(F_{\varphi_i^f}\Big(F_{\theta^s}(F_{\theta^a}(\mathbf{x})) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x}))\Big)\Big) + \max_{||\mathbf{x}'-\mathbf{x}|| \le \epsilon} \{F_{\mathrm{ce}}(\mathbf{y}, \mathcal{F}'_{\mathrm{p}}(\mathbf{x}', \mathcal{E}_j))\}\Big\}\Big\}\,, \end{aligned} \tag{6}$$

where $\mathbf{x}'$ signifies an adversarial instance of $\mathbf{x}$, synthesized via the expert $\mathcal{E}_j$, with ϵ representing the magnitude of the random vector perturbation. F_{ce} is the cross-entropy loss function defined as :

$$F_{\mathrm{ce}}(\mathbf{y}', \mathbf{y}) = \sum\nolimits_{c=1}^{C'} \left\{ \mathbf{y}[c] \log(\mathbf{y}'[c]) \right\}, \tag{7}$$

where $\mathbf{y}[c]$ and $\mathbf{y}'[c]$ denote the c-th dimension of the class label $\mathbf{y}$ and the prediction $\mathcal{F}'_{\mathrm{p}}(\mathbf{x}', \mathcal{E}_j)$, respectively. C' represents the total number of categories. To mitigate catastrophic forgetting, the dynamic backbone's update must preserve feature statistical parity across experts during novel task acquisition. To achieve this, the proposed RDRO methodology initially generates two distinct feature vector sets for a given data batch $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_b\}$, leveraging the i-th historical expert, which is constructed using previously acquired and currently learned backbone parameters, expressed as :

$$\begin{aligned} \mathbf{Z}^i &= \left\{ \mathbf{z}_t \,|\, \mathbf{z}_t = F_{\varphi_i^f}\Big(F_{\theta^d}(F_{\theta^a}(\mathbf{x}_t)) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x}_t)) \Big), t = 1, \cdots, b \right\}, \\ \hat{\mathbf{Z}}^i &= \left\{ \mathbf{z}_t \,|\, \mathbf{z}_t = F_{\varphi_i^f}\Big(F_{\theta^s}(F_{\theta^a}(\mathbf{x}_t)) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x}_t)) \Big), t = 1, \cdots, b \right\}, \end{aligned} \tag{8}$$

where b denotes the batch size. In this study, we propose to formulate $\mathbf{Z}^i$ and $\hat{\mathbf{Z}}^i$ as distributions and minimize their probabilistic divergence as a regularization loss in the primary objective function. Specifically, we propose to employ Maximum Mean Discrepancy (MMD) [31] as the distance metric, owing to its facile implementation and the robust kernel-based theoretical foundation that facilitates formal analysis. The MMD criterion serves to quantify the discrepancy between two probability density functions. This distance measure is built on the embedding of probabilities within a Reproducing Kernel Hilbert Space (RKHS) [31]. Let $P(\mathbf{Z}^i)$ and $P(\hat{\mathbf{Z}}^i)$ denote Borel probability measures for $\mathbf{Z}^i$ and $\hat{\mathbf{Z}}^i$, respectively. We consider $\mathbf{z}^i$ and $\hat{\mathbf{z}}^i$ as random variables over a topological space $\mathcal{Z}^f$. We employ $\{f \in \mathcal{F} | f \colon \mathcal{X} \to \mathbf{R}\}$ to denote a function, with $\mathcal{F}$ representing a function class. The MMD criterion between $P(\mathbf{Z}^i)$ and $P(\hat{\mathbf{Z}}^i)$ is defined as [31].

$$\mathcal{L}_{\mathrm{M}}(P(\mathbf{Z}^i), P(\hat{\mathbf{Z}}^i)) \triangleq \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathbf{z}^i \sim P(\mathbf{Z}^i)} \left[f(\mathbf{z}^i) \right] - \mathbb{E}_{\hat{\mathbf{z}}^i \sim P(\hat{\mathbf{Z}}^i)} \left[f(\hat{\mathbf{z}}^i) \right] \right). \tag{9}$$

where sup denotes the least upper bound of a set of numbers. If $P(\mathbf{Z}^i) = P(\hat{\mathbf{Z}}^i)$, we have $\mathcal{L}_{\mathrm{M}}(P(\mathbf{Z}^i), P(\hat{\mathbf{Z}}^i)) = 0$. The function class $\mathcal{F}$ is considered as a unit ball in an RKHS with a positive definite kernel $k(\mathbf{x}, \mathbf{x}')$. Calculating Eq. (9) is usually computationally intractable. In practice, the MMD is estimated on the embedding space [21], expressed as :

$$\mathcal{L}^2_{\mathrm{M}}(P(\mathbf{Z}^i), P(\hat{\mathbf{Z}}^i)) = ||\boldsymbol{\mu}_{P(\mathbf{Z}^i)} - \boldsymbol{\mu}_{P(\hat{\mathbf{Z}}^i)}||^2, \tag{10}$$

where $\boldsymbol{\mu}_{P(\mathbf{Z}^i)}$ and $\boldsymbol{\mu}_{P(\hat{\mathbf{Z}}^i)}$ denote the mean embedding of $P(\mathbf{Z}^i)$ and $P(\hat{\mathbf{Z}}^i)$, respectively. $||\cdot||^2$ denotes the Euclidean distance. $\boldsymbol{\mu}_P$ is defined as $\boldsymbol{\mu}_{P(\mathbf{Z}^i)} = \int k(\mathbf{z}^i, \cdot) \frac{\partial P(\hat{\mathbf{Z}}^i)(\mathbf{z}^i)}{\partial \mathbf{z}^i} d\mathbf{z}^i$, where $P(\hat{\mathbf{Z}}^i)(\mathbf{z}^i)$ denotes the probability density function for $P(\mathbf{Z}^i)$. $\boldsymbol{\mu}_{P(\mathbf{Z}^i)}$ also satisfies $\mathbb{E}[f(\mathbf{z}^i)] = \langle f, \boldsymbol{\mu}_{P(\mathbf{Z}^i)} \rangle_{\mathcal{H}}$, where $\langle f \ , \cdot \rangle_{\mathcal{H}}$ denotes the inner product. Since RKHS has the reproducing property $f \in \mathcal{F}, f(\mathbf{z}^i) = \langle f \ , k(\mathbf{z}^i, \cdot) \rangle_{\mathcal{H}}$, Eq. (10) can be calculated using the kernel functions, expressed as :

$$\begin{aligned} \mathcal{L}^2_{\mathrm{M}}(P(\mathbf{Z}^i), P(\hat{\mathbf{Z}}^i)) = \mathbb{E}_{\mathbf{z}^i, \mathbf{z}^{i\prime} \sim P(\mathbf{Z}^i)}[k(\mathbf{z}^i, \hat{\mathbf{z}}^i)] - 2\mathbb{E}_{\mathbf{z}^i \sim P(\mathbf{Z}^i), \hat{\mathbf{z}}^i \sim P(\hat{\mathbf{Z}}^i)}[k(\mathbf{z}^i, \hat{\mathbf{z}}^i)] \\ + \mathbb{E}_{\hat{\mathbf{z}}^i, \hat{\mathbf{z}}^{i\prime} \sim P(\hat{\mathbf{Z}}^i)}[k(\hat{\mathbf{z}}^i, \hat{\mathbf{z}}^{i'})], \end{aligned} \tag{11}$$

where $\mathbf{z}^{i'}$ and $\hat{\mathbf{z}}^{i'}$ are independent copies of $\mathbf{z}^i$ and $\hat{\mathbf{x}}^i$, respectively. In practice, we employ the same number of samples from $P(\mathbf{Z}^i)$ and $P(\hat{\mathbf{Z}}^i)$ ($N_{P(\mathbf{Z}^i)} = N_{P(\hat{\mathbf{Z}}^i)}$), where $N_{P(\mathbf{Z}^i)}$ and $N_{P(\hat{\mathbf{Z}}^i)}$ are the number of samples for $P(\mathbf{Z}^i)$ and $P(\hat{\mathbf{Z}}^i)$, respectively. Then Eq. (11) can be estimated using an unbiased empirical estimate, defined as :

$$\mathcal{L}^e_{\mathrm{M}}(P(\mathbf{Z}^i), P(\hat{\mathbf{Z}}^i)) = \frac{1}{N_{P(\mathbf{Z}^i)}(N_{P(\mathbf{Z}^i)} - 1)} \sum\nolimits_{i \neq j}^{N_{P(\mathbf{Z}^i)}} \Big\{ h(i, j) \Big\}, \tag{12}$$

where $h(i, j) = k(\mathbf{z}^i, \mathbf{z}^j) + k(\hat{\mathbf{z}}^i, \hat{\mathbf{z}}^j) - k(\mathbf{z}^i, \hat{\mathbf{z}}^j) - k(\mathbf{z}^j, \mathbf{z}^i)$. In addition, we also consider forming two groups of feature vectors using adversarial samples generated using the currently learned expert

at the j-th task learning, expressed as :

$$\begin{aligned}\mathbf{Z}^{i'} &= \left\{\mathbf{z}'_t \,|\, \mathbf{z}'_t = F_{\varphi_i^f}\Big(F_{\theta^d}(F_{\theta^a}(\mathbf{x}_t)) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x}'_t))\Big), \mathbf{x}'_t = \mathbf{x}_t + \bigtriangledown_{\mathbf{x}} F_{\mathrm{ce}}(\mathcal{F}'_{\mathrm{p}}(\mathbf{x}, \mathcal{E}_j), \mathbf{y}_t)\right\}, \\ \hat{\mathbf{Z}}^{i'} &= \left\{\mathbf{z}'_t \,|\, \mathbf{z}'_t = F_{\varphi_i^f}\Big(F_{\theta^s}(F_{\theta^a}(\mathbf{x}_t)) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x}_t))\Big), \mathbf{x}'_t = \mathbf{x}_t + \bigtriangledown_{\mathbf{x}} F_{\mathrm{ce}}(\mathcal{F}'_{\mathrm{p}}(\mathbf{x}, \mathcal{E}_j), \mathbf{y}_t)\right\},\end{aligned} \tag{13}$$

Let $P(\mathbf{Z}^{i'})$ and $P(\hat{\mathbf{Z}}^{i'})$ represent two Borel probability measures for $\mathbf{Z}^{i'}$ and $\hat{\mathbf{Z}}^{i'}$, respectively. Based on the MMD criterion, the proposed RDRO approach includes a regularization loss term for the representations at the j-th task learning, expressed as :

$$F_{\mathrm{feature}} = \min_{\theta^d} \left\{ \frac{1}{j-1} \sum\nolimits_{i=1}^{j-1} \left\{ \mathcal{L}^e_{\mathrm{M}}(P(\mathbf{Z}^i), P(\hat{\mathbf{Z}}^i)) + \mathcal{L}^e_{\mathrm{M}}(P(\mathbf{Z}^{i'}), P(\hat{\mathbf{Z}}^{i'})) \right\} \right\}. \tag{14}$$

Based on the loss terms defined in Eq. (6) and Eq. (14), the final objective function for optimizing the dynamic backbone is expressed as :

$$F_{\mathrm{RDRO}} = F_{\mathrm{feature}} + F'_{\mathrm{pre}}\,. \tag{15}$$

Furthermore, given the static backbone's immutable nature, its utilization in regulating the dynamic backbone's optimization may engender over-regularization, thereby constricting the capacity for novel task acquisition. To mitigate this, the proposed RDRO methodology effectuates a weight transfer from the dynamic backbone to the static backbone subsequent to each task transition. This design facilitates the incremental preservation of novel information within the static backbone, consequently alleviating over-regularization phenomena.

3.4 Robust Feature Fusion via Mutual Information

Many existing studies in continual learning usually utilize all active parameters to facilitate new task learning, often disregarding previously acquired representations. The utilization of critical historical representations is posited to engender positive knowledge transfer effects, thereby enhancing performance. To this end, this paper introduces a novel Mutual Information-Based Robust Feature Fusion (MBRFF) approach, which automatically ascertains knowledge similarity between each historical expert and the new task via a mutual information criterion. Specifically, during a given task learning phase ($\mathcal{T}_j$), the proposed MBRFF approach initially establishes the joint distribution $P(\mathbf{Y}^i, \mathbf{Y})$, where $P(\mathbf{Y})$ and $P(\mathbf{Y}^i)$ represent the marginal distributions of the true class labels and the corresponding predictions made using the i-th expert, respectively. Let $\mathbf{Y}^i$ and $\mathbf{Y}$ denote the random variables of the joint distribution $P(\mathbf{Y}^i, \mathbf{Y})$. The mutual information between $\mathbf{Y}^i$ and $\mathbf{Y}$ is defined as follows :

$$I(\mathbf{Y}^i; \mathbf{Y}) = \sum\nolimits_{\mathbf{y}^i \in \mathbf{Y}^i} \left\{ \sum\nolimits_{\mathbf{y} \in \mathbf{Y}} \left\{ P(\mathbf{Y}^i, \mathbf{Y})(\mathbf{y}^i, \mathbf{y}) \log \frac{P(\mathbf{Y}^i, \mathbf{Y})(\mathbf{y}^i, \mathbf{y})}{p(\mathbf{Y}^i)(\mathbf{y}^i) p(\mathbf{Y})(\mathbf{y})} \right\} \right\}, \tag{16}$$

where $P(\mathbf{Y}^i, \mathbf{Y})(\mathbf{y}^i, \mathbf{y})$ signifies the probability density function of $P(\mathbf{Y}^i, \mathbf{Y})$. The mutual information term $I(\mathbf{Y}^i; \mathbf{Y})$, as defined in Eq. (16), quantifies the degree of familiarity exhibited by the i-th expert concerning the novel task $\mathcal{T}_j$. To mitigate potential numerical overflow, the proposed MBRFF methodology normalizes the mutual information terms, subsequently employing them as adaptive weights to modulate the significance of each historical expert during the learning phase of a new task, as articulated by :

$$\alpha_i = \frac{\exp(I(\mathbf{Y}^i; \mathbf{Y}))}{\sum_{c=1}^{j-1} \{\exp(I(\mathbf{Y}^c; \mathbf{Y}))\}}\,, \tag{17}$$

where $\exp(\cdot)$ is the exponential function and α_i is the adaptive weight for the i-th expert. By utilizing Eq. (17), we can integrate representations from all history experts to form an augmented representation, expressed as :

$$\mathbf{Z}^{\mathrm{aug}} = \sum\nolimits_{i=1}^{j-1} \left\{ \alpha_i F_{\varphi_j^f}\Big(F_{\gamma_i}(\mathbf{x})[0] F_{\theta^d}(F_{\theta^a}(\mathbf{x})) \otimes F_{\gamma_i}(\mathbf{x})[1] F_{\theta^s}(F_{\theta^a}(\mathbf{x}))\Big) \right\}. \tag{18}$$

Based on the augmented representations defined in Eq. (18), the prediction process of the j-th expert can be expressed as :

$$\mathcal{F}'_{\mathrm{aug}}(\mathbf{x}, \mathcal{E}_j) = F_{\varphi_j^c}\Big(\mathbf{Z}^{\mathrm{aug}} \otimes F_{\varphi_j^f}\Big(F_{\gamma_j}(\mathbf{x})[0] F_{\theta^d}(F_{\theta^a}(\mathbf{x})) \otimes F_{\gamma_j}(\mathbf{x})[1] F_{\theta^s}(F_{\theta^a}(\mathbf{x}))\Big)\Big). \tag{19}$$

Compared to Eq. (4), the prediction process defined in Eq. (19) involves all previously learned robust representations and thus can achieve robust predictions. The pseudocode can be found in **Appendix-B** from the Supplementary Material (SM).

Split CIFAR-10								
Methods	Refresh	Refresh (Adv)	DER	DER(Adv)	DER++	DER++ (Adv)	AIR	DSEF
Clean	**92.47%**	91.76%	92.46%	91.49%	91.42%	91.70%	49.80%	90.72%
FGSM	55.84%	58.09%	55.51%	60.78%	56.13%	42.39%	18.64%	**82.36**%
PGD	05.32%	06.43%	05.79%	07.29%	05.29%	06.23%	03.92%	**79.79**%
PGDL2	65.87%	68.64%	64.42%	69.58%	64.28%	52.17%	22.34%	**82.43**%
BIM	48.69%	47.96%	50.60%	48.79%	47.63%	48.47%	16.65%	**87.43**%
CW	00.39%	00.34%	00.39%	00.19%	00.27%	00.76%	00.29%	**82.43**%
AutoAttack	03.17%	04.79%	02.06%	02.77%	02.14%	03.87%	00.76%	**90.90**%
Average	38.82%	39.71%	40.12%	41.77%	38.16%	35.08%	16.05%	**85.15**%
Split CIFAR-100								
Methods	Refresh	Refresh(Adv)	DER	DER(Adv)	DER++	DER++ (Adv)	AIR	DSEF
Clean	62.24%	61.37%	52.79%	48.67%	57.74%	57.49%	23.79%	**68.17**%
FGSM	25.89%	27.42%	21.49%	21.09%	22.95%	18.74%	09.43%	**51.71**%
PGD	03.29%	04.94%	03.76%	05.27%	04.16%	04.32%	01.46%	**44.79**%
PGDL2	32.96%	34.47%	27.68%	25.74%	30.73%	24.49%	12.93%	**53.42**%
BIM	25.47%	24.17%	22.49%	21.36%	22.98%	23.18%	10.27%	**60.79**%
CW	00.58%	00.29%	00.59%	00.94%	00.56%	00.79%	00.31%	**52.68**%
AutoAttack	02.34%	03.28%	02.44%	03.73%	02.84%	03.07%	00.89%	**66.52**%
Average	21.82%	22.27%	18.74%	18.11%	20.28%	18.86%	08.44%	**56.86**%

Table 1: The classification accuracy of the standard datasets under clean and adversarial conditions.

4 Algorithm Implementation

The comprehensive learning pipeline of our proposed method is illustrated in Fig. 1. The overall procedure can be decomposed into four key steps:

Step 1: Model expansion process During the learning of the first task, we construct a shared backbone F_{θ^a}, which serves as the foundation for expert module creation. In addition, we initialize a dynamic backbone F_{θ^d} and a static backbone F_{θ^s}, which together constitute a Siamese network architecture. For each subsequent task $\mathcal{C}_i$, a new expert module $\mathcal{E}_i$ is dynamically instantiated to accommodate task-specific knowledge.

Step 2: Calculate robust optimization loss We begin by obtaining data samples from the current task, which are first processed by the foundational backbone to extract initial representations. These representations are then forwarded through both the dynamic and static backbones, resulting in corresponding feature vectors and predictions. To enhance robustness, we compute the optimization loss terms using Eq. 6 and Eq. 14, which guide the learning of both prediction accuracy and feature consistency.

Step 3: Mutual information fusion To enhance the predictive capacity of the current expert, we additionally compute the outputs of historical experts and evaluate their relevance using mutual information. The importance weights derived from this process are used to guide the aggregation, as formalized in Eq. 19.

Step 4: Optimizing the model's parameters. The primary objective function for training the i-th expert at the i-th task learning, involves the RDRO loss terms, expressed as :

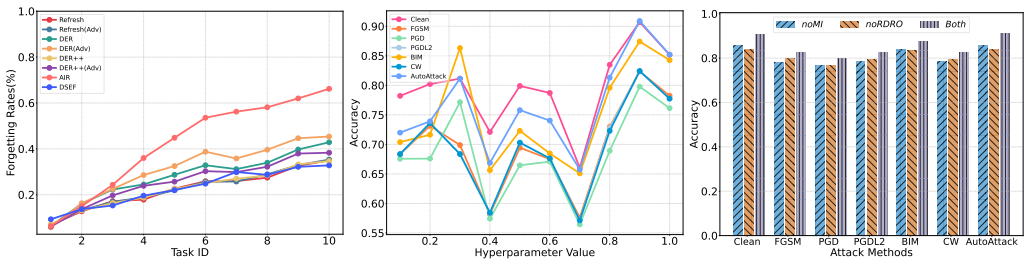
$$F(\mathbf{x}, \mathbf{y}, i) = \min_{\theta^a, \theta^s, \theta^d} \Big\{ \sum\nolimits_{c=1}^{b} \{F_{\mathrm{p}}(\mathbf{x}_c, \mathbf{y}_c) + F_{\mathrm{p}}(\mathbf{x}'_c, \mathbf{y}_c)\} + \lambda F_{\mathrm{RDRO}} \Big\}, \tag{20}$$

where $F_{\mathrm{p}}(\mathbf{x}_c, \mathbf{y}_c)$ represents the evaluation function that compares the prediction obtained from Eq. 4 with the ground-truth label $\mathbf{y}$, and $\mathbf{x}'_c$ means the adversarial sample. F_{RDRO} is defined in Eq. 15, and λ is the hyperparameter.

5 Experiment

5.1 Experimental Setting

Baselines: In this section, we present a thorough comparison between our proposed method and several established continual learning baselines, with a primary focus on experience replay-based



(a) The forgetting curve. (b) The hyperparameter λ analysis. (c) Different configurations.

Figure 2: (a) The comparison of the forgetting curves between DSEF and other baseline methods after learning a sequence of tasks. (b) The model's performance when varying λ from Eq. (20). (c) The performance of the proposed DSEF with different configurations.

Split CUB200								
Methods	Refresh	Refresh (Adv)	DER	DER(Adv)	DER++	DER++ (Adv)	AIR	DSEF
Clean	**67.62**%	61.43%	58.75%	46.55%	65.03%	53.47%	29.37%	58.47%
FGSM	26.84%	**28.52**%	21.84%	19.83%	25.18%	20.82%	10.48%	26.79%
PGD	00.46%	00.96%	00.47%	00.56%	00.42%	00.52%	00.21%	**19.83**%
PGDL2	37.85%	**39.76**%	32.17%	27.37%	35.28%	28.94%	15.58%	27.48%
BIM	22.17%	18.74%	18.16%	15.83%	21.74%	17.12%	08.33%	**34.85**%
CW	05.16%	03.56%	04.32%	05.72%	04.76%	04.25%	04.15%	**26.23**%
AutoAttack	00.21%	00.15%	00.21%	00.29%	00.17%	01.65%	09.74%	**40.74**%
Average	22.90%	21.87%	19.41%	16.59%	21.79%	18.11%	11.12%	**33.48**%
Split TinyImageNet								
Methods	Refresh	Refresh(Adv)	DER	DER(Adv)	DER++	DER++ (Adv)	AIR	DSEF
Clean	63.28%	62.36%	54.32%	52.62%	**63.36%**	60.26%	30.27%	60.21%
FGSM	25.84%	25.17%	21.68%	18.97%	26.42%	18.47%	11.75%	**45.34**%
PGD	02.38%	03.12%	01.97%	02.65%	02.46%	02.13%	00.82%	**40.13**%
PGDL2	33.78%	34.58%	31.34%	28.18%	33.48%	25.17%	13.47%	**47.78**%
BIM	22.46%	22.84%	18.57%	19.67%	23.52%	19.42%	10.94%	**55.73**%
CW	00.64%	00.42%	00.65%	00.57%	00.69%	00.67%	00.34%	**47.77**%
AutoAttack	01.12%	01.25%	00.82%	01.14%	00.94%	00.86%	00.34%	**60.80**%
Average	21.35%	21.39%	18.47%	17.68%	21.55%	18.14%	09.70%	**51.10**%

Table 2: The classification accuracy of the complex datasets under clean and adversarial conditions.

approaches. The methods evaluated include Refresh [13], DER, and DER++ [3], all of which utilize a fixed backbone throughout the training process. Since our framework incorporates adversarial training, we additionally assess the adversarial variants of these baselines, namely Refresh (Adv), DER (Adv), and DER++ (Adv), to evaluate their performance under adversarial scenarios. We also include AIR [39], a recent method designed specifically for continual adversarial defense, which treats each new class as an independent task. For a fair comparison, all replay-based methods are configured with an identical memory buffer size of 500 samples. Further details on the experimental setup can be found in **Appendix-C** from the Supplementary Material (SM).

Metrics: To systematically compare the effectiveness of different continual learning methods under adversarial settings, we adopt classification accuracy as the core performance metric across a range of training environments. After gathering all experimental outcomes, we compute an overall average accuracy for each method by aggregating results across multiple attack scenarios. To provide a comprehensive assessment of model robustness, we consider a total of seven adversarial attack strategies. These include the Fast Gradient Sign Method (FGSM) [12], Projected Gradient Descent (PGD) [24], PGD with L_2 norm, the Basic Iterative Method (BIM) [20], the Carlini and Wagner attack (CW) [4], and AutoAttack [7], which integrates several strong attacks in an ensemble fashion.

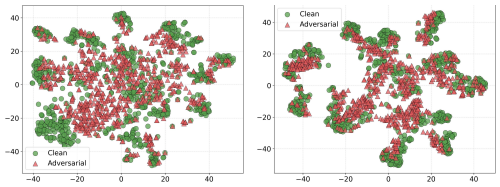
5.2 Evaluation on Standard Datasets

Table 1 presents the classification results on Split CIFAR-10 and Split CIFAR-100, comparing our proposed method with a range of state-of-the-art continual learning techniques. The empirical evidence indicates that our approach consistently delivers superior performance across both

datasets, exhibiting stronger robustness against most adversarial attack methods. While the clean accuracy of our model may be marginally lower than that of certain baselines that do not incorporate adversarial defense, our method significantly outperforms those relying on adversarial training strategies. This suggests that our framework achieves a favorable trade-off between maintaining accuracy on clean data and enhancing robustness under adversarial conditions. Notably, our method attains the highest average accuracy when considering both clean and adversarial samples, further underscoring its effectiveness in balancing standard performance and security.

5.3 Evaluation on Complex Datasets

To further assess the generalizability and robustness of different methods, we conduct experiments on more challenging benchmarks, namely Split CUB200 and Split TinyImageNet. The results are summarized in Table 2. On Split CUB200, although our method performs slightly below certain baselines in a few specific cases, it ultimately achieves the highest overall performance in terms of the average accuracy metric. This highlights its ability to maintain stable performance across diverse conditions. For the Split TinyImageNet dataset, our approach consistently outperforms all competing methods across all evaluated settings, including the average score, demonstrating its strong adaptability and effectiveness in more complex continual learning scenarios.



(a) TinyImageNet (b) CIFAR-100

Figure 3: t-SNE visualization of clean vs. adversarial samples.

5.4 Analysis Study

t-SNE Visualization. In our DSEF framework, the shared backbone is utilized to extract deep features from both clean and adversarial inputs. To better understand the impact of adversarial perturbations on feature representations, we employ t-SNE for dimensionality reduction and visualization. As illustrated in Fig. 3, the resulting embeddings show that clean and adversarial examples are closely clustered and largely overlapping in the feature space. This suggests that the backbone network is capable of mapping both types of samples into a consistent and robust representation space. Such behavior is driven by the proposed RDRO mechanism, which explicitly promotes representation invariance across different input domains. As a result, expert modules that operate on top of the shared backbone can establish more reliable decision boundaries, ultimately enhancing the model's robustness and classification accuracy. More ablation results can be found in **Appendix-D** from Supplementary Material (SM).

6 Conclusion and Limitation

This paper introduces a novel DSEF framework to enhance robustness in online continual learning by integrating a Siamese backbone with static and dynamic components. A Robust Dynamic Representation Optimization (RDRO) method is proposed to regulate dynamic updates while preserving prior knowledge. Additionally, a Mutual Information-Based Robust Feature Fusion (MBRFF) is proposed to adaptively reuse historical expert knowledge. Experiments on various benchmarks demonstrate that DSEF achieves superior performance under both clean and adversarial conditions, showcasing its effectiveness in addressing forgetting and robustness simultaneously. The primary limitation of this paper is that we adopt several popular adversarial attack methods in the experiment. In our future study, we will explore more recent adversarial attack methods to evaluate the model's performance.

7 Acknowledgments and Disclosure of Funding

This study is supported by grants from the National Natural Science Foundation of China (Grant No. 62506067, No. 62306066), the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2025XJ024, No. ZYGX2025XJ025) and Sichuan Provincial Natural Science Foundation Project (No.2025ZNSFSC0510).

References

[1] Alexander Bagnall, Razvan Bunescu, and Gordon Stewart. Training ensembles to detect adversarial examples. *arXiv preprint arXiv:1712.04006*, 2017.

[2] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8218–8227, 2021.

[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] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *In 2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[5] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. Dokania, P. H. S. Torr, and M.'A. Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

[6] 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.

[7] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020.

[8] 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.

[9] 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.

[10] 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.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, and Chang D Yoo. Forget-free continual learning with winning subnetworks. In *International Conference on Machine Learning*, pages 10734–10750. PMLR, 2022.

[19] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Proceedings of the International Conference on Learning Representations (ICLR) Workshop Track*, 2017.

[21] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

[22] K. Lin, H. F. Yang, J. H. Hsiao, and C. S. Chen. Deep learning of binary hash codes for fast image retrieval. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pages 27–35, 2015.

[23] Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017.

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[25] James Martens and Roger B. Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2408–2417. JMLR.org, 2015.

[26] 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.

[27] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[28] 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.

[29] 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.

[30] 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.

[31] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29:1930–1938, 2016.

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

[33] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019.

[34] Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. In *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:2002.06715*, 2020.

[35] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

[36] 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.

[37] F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *Proc. of Int. Conf. on Machine Learning, vol. PLMR 70*, pages 3987–3995, 2017.

[38] Guanyu Zhou, Kihyuk Sohn, and Honglak Lee. Online incremental feature learning with denoising autoencoders. In *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS), vol. PMLR 22*, pages 1453–1461, 2012.

[39] 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.

[40] 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 3

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Not included.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix-C

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We now place the codes in supplementary materials. Once this paper is accepted, we will upload it to GitHub.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 Appendix-C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Not included.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix-C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: I have checked it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Not included.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Not included.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use a standard open-source Python environment for our experiments.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have included them in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: No crowdsource.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not included.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Not include.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.