Supplementary Materials: Adversarial Robust Memory-Based Continual Learner

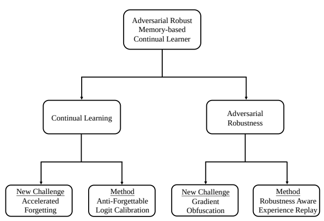

Figure S1. The mind map for the proposed adversarial robust memory-based continual learner.

In the manuscript, we propose the problem of adversarial robust continual learning. Fig. S1 shows the overall mind map of the work, and Algorithm 1 shows the full pipeline of our work. Below, we present more details and further discussions.

Algorithm 1: Memory-based Adversarial Robust Continual Learner

Input: Model M with parameters θ, memory $\mathcal{M}$, batch data $(X_1, Y_1), ..., (X_T, Y_T)$ respectively from different task distributions $\{\mathcal{D}_1, ..., \mathcal{D}_T\}$, step size of adversarial perturbations ϵ, the number of task t epochs $epoch_t$

Result: Final model M_{θ_T}

```
M ← {};
Random initialize θ_0;
for t = 1, ..., T do
    θ_t ← θ_{t-1};
    for m = 1, ..., epoch_t do
        Sampling a random batch (X_t, Y_t) ~ D_t;
        if M ≠ {} then
            Sampling a random batch (X_M, Y_M) ~ M;
            X_t ← [X_t, X_M], Y_t ← [Y_t, Y_M];
        end
        X̃_t, K_t ← PGD(θ_t, X_t, Y_t, ε);
        h_{θ_t}(X̃_t) ← M_{θ_t}(X̃_t);
        h^{lc}_{θ_t}(X̃_t) ← AFLC(h_{θ_t}(X̃_t)) Eq. (??);
        θ_t^{m+1} ← Update (θ_t^m, h^{lc}_{θ_t}(X̃_t), Y_T) Eq. (??);
        M_t ← RAER(M, X_t, K_t);
    end
    M ← M ∪ M_t;
end
```

A. Detailed Related Works

Continual learning. Continual learning [35] tries to make the model adapted to the changed data distribution following time while containing the knowledge of past data. For the main challenge: catastrophic forgetting [30, 36], existing methods can be divided into three categories: memory-based [7, 8, 37], regularization-based [21, 25, 26], and dynamic architecture [14, 39, 46, 50]. In this paper, we choose two classic settings in continual learning [29]: class incremental (class-il, models un-know task id) and task incremental (task-il, models know task id). Memory-based continual learner shows superior performance among them in either class-il or task-il settings without expanding model size [7, 56]. Hence, our research is focused on robust continual learners based on it.

What's more, some works [16, 48, 52] focus on the robustness of the continual learning under varying experimental conditions, such as task-order, memory constraints, compute constraints or time constraints, and others focus on the robustness of continual learning models against backdoor attacks [44] or privacy preservation [18]. Several recent studies [17, 19, 20] have identified the vulnerability of continual learning models to adversarial attacks, meanwhile applying adversarial sample techniques to stored data can mitigate catastrophic forgetting in continual learning, as observed in recent studies [23, 45]. Chen *et al.* [10] first tries to enhance the adversarial robustness of continual learning models by combining LwF [25] with adversarial training and using lots of unlabeled data. However, Differing from [10], we conduct an in-depth analysis of the main challenges in achieving adversarial robustness in continual learning and propose a solution that does not require additional data.

Adversarial defense. Deep neural networks usually are vulnerable to adversarial examples [15, 43, 55]. There are abundant adversarial defense methods to improve the models' adversarial robustness [1, 5]. Adversarial training has been proven the most effective way among them [12, 13, 24, 27, 28, 33, 34, 53], which leverages adversarial examples as training data. Nowadays, adversarial robustness papers mainly focus on the ideal experimental setting, Wu *et al*. [49] take into account that data in the real world often have long-tailed distributions, and Shao *et al*. [40] puts the problem of adversarial robustness in the open world. Moreover, some studies [11, 38] show that continual algorithms can facilitate rapid model adaptation to new attack methods. However, most of them are designed for single-task learning scenarios, and their effectiveness in continual learning scenarios remains largely unexplored. Different from prior works, we delve into how to improve adversarial robustness in class and task incremental settings.

B. Detailed Experimental Settings

B.1. Main Experiments

Datasets. Following common adversarial training and continual learning works [6, 34, 57], we conduct systematic analytical experiments on the Split-CIFAR10 [22] dataset and validate our improvements on the Split-CIFAR10, Split-CIFAR100 [22], and Split-Tiny-ImageNet [42] datasets. The Split-CIFAR10 contains ten classes, with $5,000$ training samples and $1,000$ test samples per class. Split-CIFAR100 consists of 100 classes, each with a set of 500 training samples and a test set of 100 samples. In the continual learning setting, Split-CIFAR10 is divided into five binary classification tasks, and Split-CIFAR100 is divided into ten tasks, each consisting of a ten-way classification task. The Split-Tiny-ImageNet has 200 classes, with 500 samples per class for training and 50 samples for validation and testing, respectively, and is split into ten tasks, where each task is a 20-way classification task.

Training details. Following common adversarial training settings, we set perturbations range of $8/255$ and step size of $2/255$ while generating adversarial samples. In the training phase, following DER and X-DER, the learning rate is 0.1, and the model architecture is ResNet18. For Split-CIFAR10, Split-CIFAR100, and Split-Tiny-ImageNet, we perform random cropping with padding of four pixels and horizontal flipping for both the stream and buffer examples. Here, we only consider how to achieve adversarial robust continual learners without large amounts of unlabeled data[1]. We train all networks using the SGD optimizer, which is also consistent with DER. The results of all experiments are run three times on different random seeds in Split-CIFAR10 and Split-CIFAR100 datasets and two times in Split-Tiny-ImageNet, and the mean and standard deviation are calculated. Due to the size of the table, we only show the mean results in the paper and put the results with standard deviations in the Appendix.

Evaluation metrics. To better measure the concerns of both adversarial robustness and continual learning, we computed Final Average Accuracy (FAA) and forgetting for the adversarial and clean samples, respectively. Projected gradient descent (PGD) attack and Auto Attack (AA) [12] are two common and effective adversarial attack methods in evaluating adversarial robustness [33, 34, 47]. PGD is a strong and classic white-box attack and we set its iteration as 20 during testing. AA is an ensemble of four diverse black-box and white-box attacks to reliably evaluate robustness, which has been proven to be reliable in evaluating deterministic defenses like adversarial training. Additional black-box attack (RayS [9]) evaluation results are in the Appendix.

Here, we set a_i^t as the accuracy for the $i-th$ task after training on task t. FAA can be defined as:

$$\text{FAA} \triangleq \frac{1}{T}\sum_{i=1}^{T} a_i^T, \tag{1}$$

and forgetting can be defined as:

$$\text{forgetting} \triangleq \frac{1}{T-1}\sum_{j=1}^{T-1} f_j,\ s.t.\ f_j = \max_{l\in\{1,\dots,T-1\}} a_i^l - a_j^T. \tag{2}$$

Forgetting ranges from [-100, 100] and measures the average decrease in accuracy, *i.e.*, the maximum difference in performance with respect to a given task observed over training.

Furthermore, CRD, FRI, and RRD in the analysis section can be defined as:

$$\text{CRD} \triangleq \text{FAA}_{\text{clean}} - \widetilde{\text{FAA}}_{\text{clean}}, \tag{3}$$

$$\text{FRI} \triangleq \widetilde{\text{Forgetting}}_{\text{clean}} - \text{Forgetting}_{\text{clean}}, \tag{4}$$

$$\text{RRD} \triangleq (\widetilde{\text{FAA}}_{\text{adv}}^{\text{Joint}} - \text{FAA}_{\text{adv}}^{\text{Joint}}) - (\widetilde{\text{FAA}}_{\text{adv}} - \text{FAA}_{\text{adv}}), \tag{5}$$

where $\text{FAA}_{\text{clean}}$ is clean data FAA of standard continual learner, $\widetilde{\text{FAA}}_{\text{clean}}$ is clean data FAA of adversarial robust continual learner; samely FAA_{adv} and $\widetilde{\text{FAA}}_{\text{adv}}$ are adversarial data FAA of standard continual learner and adversarial robust continual learner, respectively; $\text{Forgetting}_{\text{clean}}$ and $\widetilde{\text{Forgetting}}_{\text{clean}}$ are clean data forgetting of standard continual learner and adversarial robust continual learner, respectively. $\widetilde{\text{FAA}}_{\text{adv}}^{\text{Joint}}$ is the adversarial FAA of the joint adversarial learner, and $\text{FAA}_{\text{adv}}^{\text{Joint}}$ is the adversarial FAA of joint learner without adversarial training.

[1]The main external source for Chen *et al*. [10] is an 80M-TinyImage dataset, which has been withdrawn due to privacy violations.

B.2. Hyper-parameters in Analysis Experiment

When the continual algorithms are combined with Vanilla AT (AT), the input of its loss function only changes from clean samples to adversarial samples, so it will not be explained in detail.

- **ER+AT** We set the learning rate as 0.1, batch size as 32, and the number of epochs per task as 50.
- **DER+AT.** We set the learning rate as 0.03, batch size as 32, the number of epochs per task as 50, and the α in DER as 0.3.
- **DER+++AT.** We set the learning rate as 0.03, batch size as 32, the number of epochs per task as 50, and the α in DER as 0.3. The β in DER++ is 0.5 when the buffer size is 200 for Split-CIFAR10 and 500,200 for Split-CIFAR100. When the buffer size is 5120 for Split-CIFAR10, the β in DER++ is 1.0.
- **X-DER+AT.** We set the learning rate as 0.03, batch size as 32, m as 0.7, alpha is 0.05, beta is 0.01, gamma as 0.85, lambd as 0.05, eta as 0.001, temperature as 5, batch size of SimCLR loss as 64, the number of augmentation in SimCLR loss as 2, and the number of epochs per task as 10.

B.3. Baselines

Due to the particularity of our task, our baselines comprise continual learning methods and adversarial training methods, e.g. "ER+AT". For the part of continual learning baselines, we choose four popular continual learning algorithms, ER [37], DER [7], DER++ [7], and X-DER [6] in the analysis section. Furthermore, we combine our approach with two data selection-based continual learning methods: GSS [2] and ASER [41], and a logit masking-based method X-DER to show our performance in the main results section. ER randomly stores samples of past tasks and replays them in new tasks, achieving superior results without other operations; DER and DER++ store logits of old data based on ER, further alleviating catastrophic forgetting by distilling knowledge from past tasks, and DER++ additionally utilizes labels of past data to be resistant to forgetting; and X-DER embraces memory update and future preparation and uses logit masking, a special case of our AFLC, to reduce overweighting negative gradients of current data for past data. GSS selects diverse samples based on gradients. While ASER, also based on ER, utilizes the Shapley value to identify the most helpful data for mitigating forgetting.

For the part of adversarial robustness baselines, we choose four popular adversarial training algorithms: Vanilla AT [27] (abbreviated as AT in our experiments), TRADES [53], FAT [54], LBGAT [13], and SCORE [34]. AT adds the adversarial sample directly as training data, while TRADES adds a regular term that requires the adversarial sample to be consistent with the corresponding clean sample in logit outputs, both of which are currently strong robust baselines [33]. FAT chooses the adversarial sample that just succeeds in each attack to reduce clean accuracy decline in adversarial training. LBGAT achieves both robustness and clean accuracy improvements by distilling the logit of the standard training model. SCORE employs local equivariance to describe the ideal robust model's behavior to achieve top-rank performance in both robust and clean data.

Given the expensive computation of exhaustively exploring permutations of various continual learning and adversarial training algorithms, we adopt ER as the foundational baseline in combination with adversarial training algorithms based on the simplicity and effectiveness of ER+AT, and choose AT and TRADES as adversarial training baselines in evaluate the effectiveness of our approach because of the superior performance of ER+AT and ER+TRADES in adversarial FAA (Table S6).

Both continual learning and adversarial training are hyper-parameter-sensitive domains. To reduce the workload of tuning parameters, we keep the hyper-parameters of the continual learning algorithm consistent with the DER code, and we keep the hyper-parameters of the adversarial training algorithm consistent with their original papers.

Combined with different continual learning methods. When the continual algorithms are combined with Vanilla AT (AT), the input of its loss function only changes from clean samples to adversarial samples, so it will not be explained in detail.

- **ER+AT.** We set the learning rate as 0.1, batch size as 32, and the number of epochs per task as 50.
- **GSS+AT.** We set the learning rate as 0.03, batch size as 32, and the number of epochs per task as 50.
- **ASER+AT.** We set the learning rate as 0.1, batch size as 32, the maximum number of samples per class for random sampling as 1.5, the number of nearest neighbors to perform ASER as 3, and the number of epochs per task as 20.
- **X-DER+AT.** We set the learning rate as 0.03, batch size as 32, m as 0.7, alpha is 0.05, beta is 0.01, gamma as 0.85, lambd as 0.05, eta as 0.001, temperature as 5, batch size of SimCLR loss as 64, the number of augmentation in SimCLR loss as 2, and the number of epochs per task as 10.

Combined with different adversarial training methods. The learning rate, batch size, and other hyper-parameters associated with the optimization algorithm are all consistent with the ER algorithm.

- **ER+TRADES.** When ER+ TRADES combines with ours, the loss of task t can be normalized as:

$$\begin{aligned}\mathcal{L}_t \triangleq\ & \mathrm{CE}(f_\theta(x_t), y_t) + \beta * \mathrm{KL}(f_\theta(x_t), f_\theta(\widetilde{x_t})) \\ & + \mathrm{CE}(f_\theta(x_{\mathcal{M}}), y_{\mathcal{M}}) + \beta * \mathrm{KL}(f_\theta(x_{\mathcal{M}}), f_\theta(\widetilde{x_{\mathcal{M}}})),\end{aligned} \tag{6}$$

where β of TRADES is 6.0.

- **ER+FAT.** When ER+ FAT combines with ours, the loss of task t can be normalized as:

$$\mathcal{L}_t \triangleq \text{CE}(f_\theta(\widetilde{x}_t), y_t) + \text{CE}(f_\theta(\widetilde{x}_\mathcal{M}), y_\mathcal{M}). \quad (7)$$

  Note that when solving the adversarial sample in the training phase, the iteration is stopped once the attack model is successful.
- **ER+LBGAT.** Here we implement LBGAT based on TRADES ($\beta = 0.0$). When ER+ LBGAT combines with ours, the loss of task t can be normalized as:

$$\begin{aligned}\mathcal{L}_t \triangleq\ & \text{CE}(f_\theta(\widetilde{x}_t), y_t) + \gamma * \text{MSE}(f_\theta^{clean}(x_t), f_\theta(\widetilde{x}_t)) \\ & + \text{CE}(f_\theta(\widetilde{x}_\mathcal{M}), y_\mathcal{M}) \\ & + \gamma * \text{MSE}(f_\theta^{clean}(x_\mathcal{M}), f_\theta(\widetilde{x}_\mathcal{M})),\end{aligned} \quad (8)$$

  γ of LBGAT is 0.1, and f_θ^{clean} is a standard continual learning model (ER on our experiments) with the model architecture of ResNet-18.
- **ER+SCORE.** Compared with ER+TRADES, it performs better on clean samples but is less adversarial robust, probably because the hyper-parameters are unsuitable for continual learning scenarios. We implement it using β as 4.0, label smoothing as 0.1, and gradient clip g as 0.

$$\begin{aligned}\mathcal{L}_t \triangleq\ & \text{MSE}\left(f_\theta(x_t), y_t\right) \\ & + \beta * \text{ReLU}(\text{MSE}(f_\theta(x_t), f_\theta(\widetilde{x}_t)) - g) \\ & + \text{MSE}\left(f_\theta(x_\mathcal{M}), y_\mathcal{M}\right) \\ & + \beta * \text{ReLU}(\text{MSE}(f_\theta(x_\mathcal{M}), f_\theta(\widetilde{x}_\mathcal{M})) - g).\end{aligned} \quad (9)$$

- **ER+TRADES+ours.** When ER+TRADES combines with ours, the loss of task t is:

$$\begin{aligned}\mathcal{L}_t \triangleq\ & \text{CE}(f_\theta^{lc}(x_t), y_t) + \beta * \text{KL}(f_\theta^{lc}(x_t), f_\theta^{lc}(\widetilde{x}_t)) \\ & + \text{CE}(f_\theta^{lc}(x_\mathcal{M}), y_\mathcal{M}) + \beta * \text{KL}(f_\theta^{lc}(x_\mathcal{M}), f_\theta^{lc}(\widetilde{x}_\mathcal{M})),\end{aligned} \quad (10)$$

  where β of TRADES is 6.0.

C. More Experiments

In this section, we provide experiments mentioned in our paper, including: 1)Evaluations on Split-Tiny-ImageNet and more challenging datasets, 2)Extended robustness verification usingadditional black-box (RayS) and adaptive attacks; 3)Training dynamics illustrated through accuraey curves onSplit-CIFAR10 with varying buffer sizes(200,5,120); 4)Sensitivity analysis of hyperparameters (o in AFLC andp in RAER) based on ER+TRADES using Split-CIFAR10with buffer size 200. 5)Training time for different methods.

C.1. Experiments on Tiny-ImageNet

The results in Table S8 clearly demonstrate that our proposed method is also effective at improving upon the baseline algorithms on the more challenging Split-Tiny-ImageNet dataset. Specifically, our approach led to maximum improvements in clean FAA of 3.71%, adversarial FAA of 2.50%, and alleviated forgetting by up to 4.06%.

C.2. Experiments on ViT

The results in Table S10 clearly demonstrate that our proposed method is also effective at improving upon ViT. We use a ViT-based adversarial training method [31] and ER as a baseline on Split-CIFAR10. We achieve a max 33.56% clean forgetting reduction, 20.10% robust forgetting reduction, and 13.6% FAA improvement.

C.3. Ablation Experiments

As shown in Table S4, we have performed ablation experiments based on ER+TRADES under the Split-CIFAR10 dataset. The results demonstrate that AFLC (Sec. 4.2) can effectively mitigate the increased forgetting caused by adversarial training under class incremental setting (55.49% for clean samples and 43.18% for adversarial samples forgetting, with corresponding FAA improvements of 15.46% and 2.86%, respectively). AFLC does not show significant improvement in the task incremental setting due to excessive suppression of future task classification heads and the use of the same calibration value for classes within the same task.

RAER (Sec. 4.3) can further improve the robust accuracy of AFLC by 1.23% for class incremental setting and 2.07% for task incremental setting and reduce the robust forgetting by 12.16% and 3.47% respectively. That proves the data selected by RAER describe the overall data distribution more accurately and effectively mitigate the gradient obfuscation phenomenon.

When considering the future prior adjustment (FP in Table S4), we find that although the forgetting of the class incremental setting is higher, the FAA of both clean samples and adversarial samples has been significantly improved, and the forgetting of the task incremental setting has been further reduced, which proves that FP can reduce negative gradients to future classes and help learn new tasks.

Hyper-parameter sensitivity. We study the sensitivity of hyper-parameters α in AFLC and ρ in RAER on the basis of ER+TRADES with a dataset of Split-CIFAR10 and a buffer size of 200.

- **Impact of ρ.** The results are shown in Table S3. The value of ρ in the range of $[5, 10]$ is robust and ensures the selection of safe and diverse samples for storage, but when the parameter ρ is too small (1), the samples selected are safe but not diverse enough to improve the robustness of the model to a limited extent. We can also find that adding only RAER does not help the robustness of the class incremental setting, because the adversarial sample is too suppressed for the past category, and the ro-

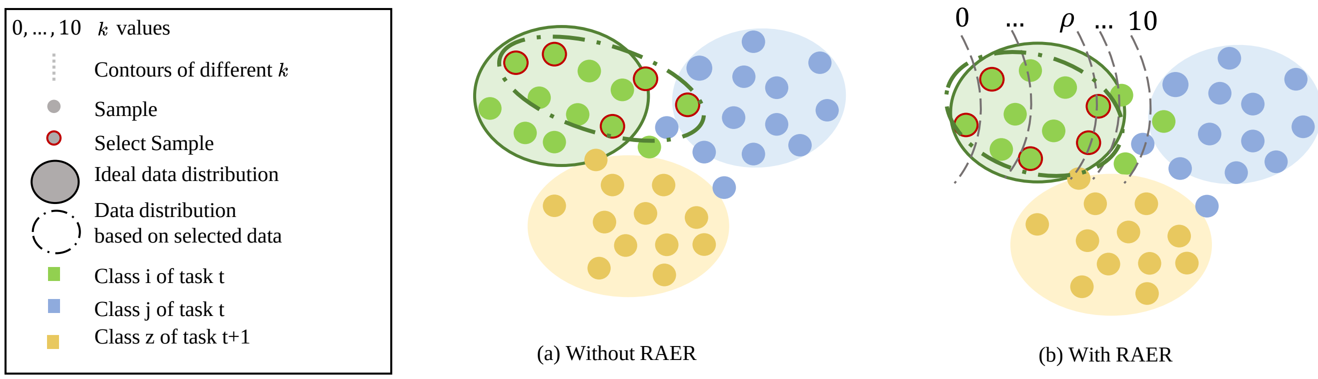

Figure S2. Schematic diagram of the RAER. A larger k means that the sample is more vulnerable, so the closer it is to its task decision boundary. RAER can exclude vulnerable samples that over-fit the boundary of the current task, thus selecting samples that are more robustly safe and more representative of the data distribution.

Methods	**Class Incremental Setting**					**Task Incremental Setting**				
	Clean Data		**Adversarial Data**			**Clean Data**		**Adversarial Data**		
	FAA↑	**Forgetting↓**	**PGD-20↑**	**AA↑**	**Forgetting↓**	**FAA↑**	**Forgetting↓**	**PGD-20↑**	**AA↑**	**Forgetting↓**
ER+AT (ViT)	$29.22_{\pm 0.62}$	$77.33_{\pm 1.01}$	$18.05_{\pm 0.12}$	$17.71_{\pm 0.42}$	$68.39_{\pm 1.23}$	$85.23_{\pm 0.62}$	$10.11_{\pm 0.86}$	$45.19_{\pm 0.99}$	$44.78_{\pm 1.21}$	$35.82_{\pm 0.98}$
ER+AT (ViT)+Ours	$\mathbf{40.41}_{\pm 0.56}$	$\mathbf{43.77}_{\pm 0.21}$	$\mathbf{30.42}_{\pm 0.11}$	$\mathbf{30.24}_{\pm 0.10}$	$\mathbf{48.29}_{\pm 0.45}$	$\mathbf{90.05}_{\pm 0.09}$	$\mathbf{7.93}_{\pm 0.07}$	$\mathbf{58.88}_{\pm 0.27}$	$\mathbf{57.62}_{\pm 0.19}$	$\mathbf{24.16}_{\pm 0.32}$

Table S1. Experiment results on Split-CIFAR10 dataset and buffer sizes of 200 in ViT. Bold represents the best experimental results for the same settings.

bustness in class incremental is improved when AFLC is added (as shown in Table S4).

- **Impact of α.** The results are shown in Table S5. Observing the experimental results, we find that as the value of α increases, the negative gradient impact of adversarial examples on the classification head of previous tasks decreases, indicating a stronger ability of the model to resist forgetting. However, when α becomes excessively large, the model's learning capacity for the current task is heavily suppressed, resulting in a decline in the model's adversarial robustness. Therefore, we choose $\alpha = 3.5$ in other experiments.

C.4. Generalization on Adaptive Attacks

As mentioned by [4, 32], generic attack methods alone are not adequate to account for solid robustness. Therefore, we use the adaptive attack based on PGD-20 and Auto Attack by doing the same logit calibration as our model training phase when generating the adversarial samples (details in supplemental materials). As shown in Table S9, we are still able to maintain stable robustness even under adaptive attacks. We find that adding logit calibration to the solution adversarial sample stage reduces the attack strength, especially for AA, and Yang *et al*. [51] also find a similar phenomenon. We conjecture that logit calibration may introduce overfitting from logit prior when generating adversarial examples.

For both PGD-20 and AA, we considered both class incremental setting and task incremental setting, and the logit in solving the adversarial sample is processed by AFLC.

$$h_\theta^{\mathrm{lc}}(\widetilde{x})_i = h_\theta(\widetilde{x})_i - \mathrm{v}_i, \tag{11}$$

where v_i is the same as the v of the last task training phase.

More Black-box Attack Method. AA used in our experiments contains a query-efficient black-box attack, Square [3]. What's more, we additionally test the robustness of our model against another strong black-box attack, RayS [9] (query limitation is $10,000$). As shown in Table S2, our method can effectively improve the robustness of the continual learners under black-box attacks.

Table S2. Defense success rate under black-box attack RayS.

Method	**S-CIFAR10 200**	**S-CIFAR10 5120**	**S-CIFAR100 500**	**S-CIFAR100 2000**
ER+TRADES	7.55	17.90	3.73	4.41
ER+TRADES+Ours	**17.49**	**26.11**	**9.61**	**11.06**

Results with Standard Error. In this section, we provide results with mean and standard error. In the task incremental setting, we observe more stable experimental results compared to the class incremental setting. Furthermore, our approach achieves consistent improvements in the majority of cases.

Accuracy during Training in Split-CIFAR10/100. Figure S3 shows FAAs of different continual training phrases of ER+AT, ER+TRADES, and their combination with us. In the vast majority of experiments, our proposed method

Table S3. Data selection strategy ablation experiments on the CIFAR-10 dataset with the buffer size of 200, using ER + TRADES as the baseline in the class/task incremental settings. When $\rho = 11$, this is equivalent to not applying the robust data selection strategy.

ρ	**Class Incremental Setting**		**Task Incremental Setting**	
	FAA↑	**PGD-20↑**	**FAA↑**	**PGD-20↑**
11	$22.42_{\pm 7.11}$	$15.72_{\pm 0.82}$	$78.79_{\pm 0.81}$	$51.33_{\pm 2.3}$
10	$18.04_{\pm 0.58}$	$15.31_{\pm 0.18}$	$80.38_{\pm 0.16}$	$59.88_{\pm 1.21}$
5	$18.21_{\pm 0.25}$	$15.50_{\pm 0.12}$	$80.52_{\pm 0.71}$	$59.50_{\pm 1.82}$
1	$18.35_{\pm 0.88}$	$15.06_{\pm 0.17}$	$79.62_{\pm 0.23}$	$56.92_{\pm 0.36}$

can improve the performance of the baseline models at each incremental training stage. Despite a minor decline in FAA on clean samples compared to baselines under the task incremental setting with a buffer size of $5,120$, this is due to the balance between robustness and performance and the forgetting alleviation from AFLC diminishes as buffer size increases. For FAA, this is attributed to that the RAER component designed for robustness improvement can adversely affect clean data performance, exacerbated by the large buffer size. The trade-off between standard and robust accuracy is an expected consequence of adversarial training, wherein improved adversarial robustness typically incurs some cost to natural sample performance. Nonetheless, our approach still confers substantial gains in adversarial robustness with limited sacrifice of conventional accuracy compared to baselines, as elucidated by the buffer size analysis on the interaction between RAER and AFLC. In addition, AFLC aims to mitigate the phenomenon of forgetting which hinders the learning of current tasks. Therefore, in the initial task, there might be a slight performance decline due to AFLC. However, the benefits of AFLC become prominent in later stages, leading to improved robustness.

Figure S4 shows FAAs of different continual training phrases of ER+AT, ER+TRADES, and their combination with us. In the vast majority of experiments, our proposed method can improve the performance of the baseline models at each incremental training stage.

C.5. Training time

Table S10 shows the GPU memory usage and training time per epoch of different methods

Table S4. Ablation experiments on the Split-CIFAR10 dataset with the buffer size of 200, using ER+TRADES as the baseline. **Bold** indicates that the inclusion of this module will relatively enhance the corresponding evaluation metrics. Experiments have demonstrated that AFLC mitigates clean-sample accelerated forgetting from adversarial samples, RAER mitigates gradient obfuscation (Adversarial FAA has a boost) and robust forgetting; adding FP improves the model's ability to learn new tasks (FAA has an overall increase).

AFLC	RAER	FP	Class Incremental Setting				Task Incremental Setting			
			Clean Data		Adversarial Data		Clean Data		Adversarial Data	
			FAA↑	Forgetting↓	FAA↑	Forgetting↓	FAA↑	Forgetting↓	FAA↑	Forgetting↓
			$22.42_{\pm7.11}$	$77.25_{\pm10.79}$	$15.72_{\pm0.82}$	$64.95_{\pm5.58}$	$78.79_{\pm0.81}$	$8.1_{\pm0.97}$	$51.33_{\pm2.3}$	$21.97_{\pm1.31}$
✓			$37.88_{\pm0.54}$	$21.76_{\pm0.23}$	$18.58_{\pm1.9}$	$21.77_{\pm0.24}$	$81.66_{\pm0.02}$	$10.86_{\pm0.78}$	$53.64_{\pm1.48}$	$17.6_{\pm0.28}$
✓	✓		$37.45_{\pm5.28}$	$\mathbf{10.29_{\pm11.36}}$	$\mathbf{19.81_{\pm1.75}}$	$\mathbf{9.61_{\pm6.38}}$	$78.13_{\pm3.1}$	$15.39_{\pm4.28}$	$\mathbf{55.71_{\pm3.47}}$	$\mathbf{14.13_{\pm3.09}}$
✓	✓	✓	$\mathbf{43.34_{\pm4.27}}$	$33.40_{\pm11.02}$	$\mathbf{19.85_{\pm1.55}}$	$30.78_{\pm8.84}$	$\mathbf{82.59_{\pm1.12}}$	$\mathbf{7.53_{\pm1.24}}$	$\mathbf{59.41_{\pm0.61}}$	$14.16_{\pm1.29}$

Table S5. Ablation experiments of different α.

α	Class Incremental Setting				Task Incremental Setting			
	Clean Data		Adversarial Data		Clean Data		Adversarial Data	
	FAA↑	Forgetting↓	FAA↑	Forgetting↓	FAA↑	Forgetting↓	FAA↑	Forgetting↓
0.0	$\mathbf{38.85_{\pm3.26}}$	$47.53_{\pm21.21}$	$17.56_{\pm1.40}$	$42.80_{\pm21.82}$	$\mathbf{82.70_{\pm0.43}}$	$\mathbf{9.97_{\pm1.04}}$	$50.67_{\pm6.55}$	$26.28_{\pm10.45}$
3.5	$37.88_{\pm0.54}$	$21.76_{\pm0.23}$	$\mathbf{18.58_{\pm1.90}}$	$21.77_{\pm0.24}$	$81.66_{\pm0.02}$	$10.86_{\pm0.78}$	$\mathbf{53.64_{\pm1.48}}$	$17.60_{\pm0.28}$
7.0	$35.92_{\pm0.88}$	$\mathbf{20.99_{\pm6.03}}$	$13.89_{\pm0.63}$	$\mathbf{21.76_{\pm0.67}}$	$82.56_{\pm0.58}$	$10.49_{\pm3.75}$	$52.69_{\pm1.18}$	$\mathbf{16.25_{\pm2.52}}$

Table S6. Experiment results on Split-CIFAR10/100 dataset and model architecture is ResNet18. Here PGD-20 and AA are adversarial data Final Average Accuracy (FAA) generated by PGD-20 and Auto Attack (AA) respectively. Forgetting of adversarial data is computed based on PGD-20. With the addition of ours, model performance can be improved across the board.

(a) Results on Split-CIFAR10. We chose two buffer sizes of 200 and 5120.

Buffer Size	Methods	Class Incremental Setting					Task Incremental Setting				
		Clean Data		Adversarial Data			Clean Data		Adversarial Data		
		FAA↑	Forgetting↓	PGD-20↑	AA↑	Forgetting↓	FAA↑	Forgetting↓	PGD-20↑	AA↑	Forgetting↓
200	ER+AT	$28.18_{\pm0.69}$	$80.58_{\pm1.05}$	$17.86_{\pm0.29}$	$16.94_{\pm0.38}$	$69.58_{\pm1.15}$	$84.49_{\pm0.61}$	$10.23_{\pm0.98}$	$44.30_{\pm1.05}$	$44.69_{\pm1.04}$	$36.89_{\pm1.11}$
	ER+TRADES	$22.42_{\pm7.11}$	$77.25_{\pm10.79}$	$15.72_{\pm0.82}$	$15.53_{\pm0.68}$	$64.95_{\pm5.58}$	$78.79_{\pm0.81}$	$8.10_{\pm0.97}$	$51.33_{\pm2.30}$	$51.50_{\pm2.33}$	$21.97_{\pm1.31}$
	ER+FAT	$33.61_{\pm6.80}$	$69.21_{\pm8.95}$	$15.14_{\pm0.80}$	$14.81_{\pm0.92}$	$49.04_{\pm7.29}$	$83.40_{\pm0.92}$	$10.35_{\pm0.83}$	$43.69_{\pm2.08}$	$43.96_{\pm2.11}$	$28.56_{\pm2.14}$
	ER+LBGAT	$25.68_{\pm0.56}$	$84.47_{\pm0.63}$	$16.65_{\pm0.11}$	$16.56_{\pm0.09}$	$70.5_{\pm0.29}$	$78.19_{\pm1.17}$	$18.85_{\pm1.58}$	$40.69_{\pm3.03}$	$40.73_{\pm3.15}$	$40.83_{\pm3.80}$
	ER+Pang *et al*. [34]	$48.65_{\pm1.76}$	$56.79_{\pm1.78}$	$2.40_{\pm0.46}$	$0.93_{\pm0.03}$	$18.96_{\pm2.47}$	$88.90_{\pm2.02}$	$9.82_{\pm2.53}$	$7.25_{\pm1.06}$	$6.72_{\pm0.55}$	$8.70_{\pm0.06}$
	ER+AT+Ours	$35.68_{\pm0.57}$	$71.18_{\pm0.66}$	$18.40_{\pm0.56}$	$18.16_{\pm0.50}$	$67.85_{\pm0.72}$	$84.87_{\pm0.56}$	$9.93_{\pm0.60}$	$47.30_{\pm1.31}$	$47.61_{\pm1.35}$	$34.04_{\pm1.42}$
	ER+TRADES+Ours	$43.34_{\pm4.27}$	$33.40_{\pm11.02}$	$19.85_{\pm1.55}$	$18.35_{\pm1.14}$	$30.78_{\pm8.84}$	$82.59_{\pm1.12}$	$7.53_{\pm1.24}$	$59.41_{\pm0.61}$	$59.59_{\pm0.64}$	$14.16_{\pm1.29}$
5120	ER+AT	$61.88_{\pm0.74}$	$37.72_{\pm0.73}$	$27.28_{\pm0.56}$	$26.69_{\pm0.52}$	$41.66_{\pm1.22}$	$91.24_{\pm0.19}$	$2.56_{\pm0.12}$	$56.59_{\pm0.88}$	$56.90_{\pm0.88}$	$19.34_{\pm1.11}$
	ER+TRADES	$20.36_{\pm2.81}$	$85.14_{\pm4.28}$	$16.3_{\pm0.44}$	$16.18_{\pm0.35}$	$72.85_{\pm1.46}$	$88.48_{\pm0.81}$	$1.59_{\pm0.86}$	$64.36_{\pm0.51}$	$64.52_{\pm0.48}$	$12.47_{\pm0.37}$
	ER+FAT	$54.55_{\pm6.71}$	$43.18_{\pm7.51}$	$19.68_{\pm2.54}$	$18.91_{\pm2.38}$	$42.15_{\pm3.59}$	$91.50_{\pm0.53}$	$2.12_{\pm0.80}$	$56.72_{\pm0.65}$	$56.87_{\pm0.70}$	$14.79_{\pm1.13}$
	ER+LBGAT	$62.45_{\pm0.46}$	$37.73_{\pm0.97}$	$27.42_{\pm0.27}$	$26.66_{\pm0.33}$	$47.83_{\pm0.49}$	$91.10_{\pm0.62}$	$3.25_{\pm1.06}$	$56.57_{\pm1.02}$	$56.31_{\pm1.13}$	$19.38_{\pm0.84}$
	ER+Pang *et al*. [34]	$51.37_{\pm6.44}$	$52.84_{\pm8.47}$	$3.14_{\pm0.49}$	$1.10_{\pm0.05}$	$20.29_{\pm2.92}$	$95.63_{\pm0.17}$	$1.58_{\pm0.21}$	$14.66_{\pm0.15}$	$13.92_{\pm0.04}$	$2.53_{\pm0.46}$
	ER+AT+Ours	$64.34_{\pm0.68}$	$23.64_{\pm0.23}$	$31.31_{\pm0.07}$	$30.49_{\pm0.06}$	$20.46_{\pm0.75}$	$91.0_{\pm0.07}$	$3.51_{\pm0.14}$	$60.49_{\pm0.24}$	$60.61_{\pm0.28}$	$13.27_{\pm0.33}$
	ER+TRADES+Ours	$39.80_{\pm12.23}$	$44.08_{\pm18.66}$	$23.07_{\pm6.10}$	$21.98_{\pm5.41}$	$41.37_{\pm11.41}$	$86.48_{\pm2.55}$	$1.71_{\pm0.62}$	$69.25_{\pm2.23}$	$69.38_{\pm2.23}$	$5.33_{\pm0.74}$

(b) Results on Split-CIFAR100. We choose two buffer sizes of 500 and 2000.

Buffer Size	Methods	Class Incremental Setting					Task Incremental Setting				
		Clean Data		Adversarial Data			Clean Data		Adversarial Data		
		FAA↑	Forgetting↓	PGD-20↑	AA↑	Forgetting↓	FAA↑	Forgetting↓	PGD-20↑	AA↑	Forgetting↓
500	ER+AT	$11.94_{\pm0.74}$	$73.54_{\pm0.75}$	$5.66_{\pm0.31}$	$5.54_{\pm0.31}$	$38.03_{\pm0.10}$	$52.71_{\pm0.96}$	$28.44_{\pm1.11}$	$17.35_{\pm0.08}$	$19.56_{\pm0.10}$	$25.67_{\pm0.26}$
	ER+TRADES	$7.59_{\pm0.08}$	$68.13_{\pm0.61}$	$5.43_{\pm0.05}$	$5.11_{\pm0.06}$	$44.13_{\pm0.89}$	$50.75_{\pm0.42}$	$20.22_{\pm0.49}$	$26.89_{\pm0.58}$	$27.69_{\pm0.56}$	$20.34_{\pm0.72}$
	ER+FAT	$11.48_{\pm0.72}$	$73.99_{\pm1.18}$	$5.35_{\pm0.26}$	$5.23_{\pm0.27}$	$37.26_{\pm0.62}$	$56.1_{\pm1.23}$	$24.71_{\pm1.00}$	$20.01_{\pm0.85}$	$22.99_{\pm0.91}$	$22.31_{\pm1.10}$
	ER+LBGAT	$11.77_{\pm0.37}$	$75.1_{\pm0.40}$	$6.16_{\pm0.11}$	$5.82_{\pm0.13}$	$39.77_{\pm0.18}$	$42.02_{\pm2.36}$	$26.45_{\pm0.30}$	$13.99_{\pm0.83}$	$14.43_{\pm0.97}$	$23.14_{\pm0.21}$
	ER+Pang *et al*. [34]	$16.22_{\pm0.63}$	$77.71_{\pm0.44}$	$0.91_{\pm0.05}$	$0.62_{\pm0.00}$	$6.13_{\pm0.16}$	$68.87_{\pm0.25}$	$11.21_{\pm0.73}$	$3.23_{\pm0.42}$	$7.95_{\pm0.06}$	$4.94_{\pm0.05}$
	ER+AT+Ours	$24.14_{\pm0.44}$	$53.03_{\pm0.3}$	$7.13_{\pm0.05}$	$6.68_{\pm0.04}$	$25.81_{\pm0.13}$	$56.0_{\pm0.05}$	$26.19_{\pm0.03}$	$17.95_{\pm0.04}$	$20.26_{\pm0.06}$	$24.87_{\pm0.08}$
	ER+TRADES+Ours	$23.24_{\pm0.01}$	$24.29_{\pm0.0}$	$9.93_{\pm0.37}$	$7.5_{\pm0.37}$	$11.7_{\pm3.32}$	$56.68_{\pm0.44}$	$21.39_{\pm0.12}$	$27.39_{\pm0.41}$	$28.5_{\pm0.34}$	$16.76_{\pm0.22}$
2000	ER+AT	$18.77_{\pm0.18}$	$65.06_{\pm0.58}$	$7.20_{\pm0.18}$	$7.01_{\pm0.18}$	$33.56_{\pm0.26}$	$62.01_{\pm0.76}$	$17.97_{\pm0.80}$	$21.16_{\pm0.38}$	$24.04_{\pm0.23}$	$20.47_{\pm0.25}$
	ER+TRADES	$9.50_{\pm0.24}$	$70.49_{\pm0.38}$	$5.35_{\pm0.19}$	$5.01_{\pm0.17}$	$42.08_{\pm0.40}$	$60.63_{\pm1.03}$	$13.78_{\pm0.90}$	$26.68_{\pm0.51}$	$29.19_{\pm0.70}$	$18.60_{\pm0.54}$
	ER+FAT	$17.09_{\pm1.96}$	$66.91_{\pm3.02}$	$6.12_{\pm0.40}$	$5.93_{\pm0.38}$	$33.62_{\pm1.72}$	$63.88_{\pm0.02}$	$15.99_{\pm0.47}$	$23.48_{\pm0.06}$	$26.75_{\pm0.05}$	$17.12_{\pm0.62}$
	ER+LBGAT	$20.58_{\pm0.15}$	$63.31_{\pm0.20}$	$7.93_{\pm0.17}$	$7.05_{\pm0.16}$	$30.09_{\pm0.32}$	$55.77_{\pm0.23}$	$25.11_{\pm0.54}$	$17.95_{\pm0.43}$	$18.88_{\pm0.54}$	$19.18_{\pm0.48}$
	ER+Pang *et al*. [34]	$30.10_{\pm0.53}$	$61.51_{\pm1.00}$	$0.75_{\pm0.22}$	$0.51_{\pm0.02}$	$4.10_{\pm0.60}$	$76.90_{\pm0.44}$	$19.60_{\pm0.32}$	$4.88_{\pm0.24}$	$11.97_{\pm0.26}$	$5.25_{\pm0.20}$
	ER+AT+Ours	$31.93_{\pm0.07}$	$40.5_{\pm0.12}$	$9.61_{\pm0.06}$	$9.16_{\pm0.04}$	$17.78_{\pm0.22}$	$63.77_{\pm0.18}$	$17.16_{\pm0.31}$	$23.11_{\pm0.02}$	$25.59_{\pm0.08}$	$17.57_{\pm0.07}$
	ER+TRADES+Ours	$28.73_{\pm1.79}$	$24.16_{\pm2.6}$	$12.75_{\pm0.34}$	$11.02_{\pm0.66}$	$14.56_{\pm0.7}$	$62.01_{\pm0.89}$	$13.16_{\pm1.9}$	$34.81_{\pm0.41}$	$35.36_{\pm0.95}$	$11.6_{\pm0.17}$

Table S7. Experiments with other data selection-based and logit masking-based continual learning methods.

Method	Publication	Class Incremental Setting				Task Incremental Setting			
		Clean Data		Adversarial Data		Clean Data		Adversarial Data	
		FAA↑	**Forgetting↓**	**FAA↑**	**Forgetting↓**	**FAA↑**	**Forgetting↓**	**FAA↑**	**Forgetting↓**
ER+AT	NeurIPS 2019	$28.18_{\pm 0.69}$	$80.58_{\pm 1.05}$	$17.86_{\pm 0.29}$	$69.58_{\pm 1.15}$	$84.49_{\pm 0.61}$	$10.23_{\pm 0.98}$	$44.30_{\pm 1.05}$	$36.89_{\pm 1.11}$
ER+AT+Ours		$47.70_{\pm 0.67}$	$53.19_{\pm 1.87}$	$18.20_{\pm 0.02}$	$56.73_{\pm 1.10}$	$85.40_{\pm 1.28}$	$9.20_{\pm 1.67}$	$48.38_{\pm 0.67}$	$31.68_{\pm 1.49}$
GSS+AT	NeurIPS 2019	$27.59_{\pm 0.62}$	$80.78_{\pm 0.99}$	$16.67_{\pm 0.16}$	$68.53_{\pm 0.11}$	$84.41_{\pm 0.10}$	$9.83_{\pm 0.13}$	$44.25_{\pm 0.87}$	$34.79_{\pm 0.09}$
GSS+AT+Ours		$36.93_{\pm 7.13}$	$67.72_{\pm 15.65}$	$16.84_{\pm 0.01}$	$60.04_{\pm 16.20}$	$85.57_{\pm 0.56}$	$8.86_{\pm 1.84}$	$47.11_{\pm 0.04}$	$31.83_{\pm 0.36}$
ASER+AT	AAAI 2021	$18.85_{\pm 0.00}$	$87.78_{\pm 0.34}$	$14.06_{\pm 0.57}$	$65.57_{\pm 0.01}$	$73.87_{\pm 6.23}$	$19.01_{\pm 13.46}$	$30.65_{\pm 14.44}$	$44.85_{\pm 30.80}$
ASER+AT+Ours		$24.45_{\pm 0.03}$	$81.73_{\pm 0.60}$	$14.91_{\pm 0.07}$	$62.74_{\pm 0.66}$	$77.70_{\pm 0.08}$	$15.21_{\pm 0.43}$	$34.50_{\pm 0.15}$	$38.55_{\pm 0.39}$
X-DER+AT	TPAMI 2022	$34.04_{\pm 0.81}$	$25.13_{\pm 16.25}$	$16.82_{\pm 0.95}$	$27.84_{\pm 17.33}$	$80.80_{\pm 0.44}$	$4.96_{\pm 2.74}$	$60.83_{\pm 0.31}$	$10.99_{\pm 0.65}$
X-DER+AT+Ours		$43.25_{\pm 0.09}$	$20.77_{\pm 2.56}$	$17.22_{\pm 0.00}$	$18.56_{\pm 11.86}$	$84.87_{\pm 0.00}$	$1.74_{\pm 0.02}$	$61.68_{\pm 0.00}$	$7.03_{\pm 1.50}$

Table S8. Experiment results on S-Tiny-ImageNet dataset and model architecture is ResNet18. Following [7], we choose two buffer sizes of 200 and 5120. Here PGD-20 and AA are adversarial data Final Average Accuracy (FAA) generated by PGD-20 and Auto Attack (AA) respectively. Forgetting of adversarial data is computed based on PGD-20. **Bold** represents the best experimental results for the same settings. With the addition of ours, model performance can be improved across the board.

Buffer Size	Methods	Class Incremental Setting					Task Incremental Setting				
		Clean Data		Adversarial Data			Clean Data		Adversarial Data		
		FAA↑	**Forgetting↓**	**PGD-20↑**	**AA↑**	**Forgetting↓**	**FAA↑**	**Forgetting↓**	**PGD-20↑**	**AA↑**	**Forgetting↓**
200	**ER+TRADES**	$5.50_{\pm 0.41}$	$54.67_{\pm 0.13}$	$2.04_{\pm 0.35}$	$1.92_{\pm 0.35}$	$22.62_{\pm 0.21}$	$23.51_{\pm 1.38}$	$34.72_{\pm 1.15}$	$5.28_{\pm 0.73}$	$6.48_{\pm 0.91}$	$19.10_{\pm 0.42}$
	ER+TRADES+Ours	$\mathbf{7.35_{\pm 0.43}}$	$\mathbf{52.38_{\pm 0.22}}$	$\mathbf{2.22_{\pm 0.08}}$	$\mathbf{2.05_{\pm 0.03}}$	$\mathbf{20.79_{\pm 0.14}}$	$\mathbf{25.58_{\pm 0.73}}$	$\mathbf{34.94_{\pm 1.44}}$	$\mathbf{6.24_{\pm 0.30}}$	$\mathbf{7.60_{\pm 0.26}}$	$\mathbf{18.41_{\pm 0.20}}$
5120	**ER+TRADES**	$7.24_{\pm 0.39}$	$55.69_{\pm 1.06}$	$2.39_{\pm 0.16}$	$2.21_{\pm 0.14}$	$21.59_{\pm 0.91}$	$41.36_{\pm 1.07}$	$18.16_{\pm 1.74}$	$11.01_{\pm 0.92}$	$13.71_{\pm 0.78}$	$12.61_{\pm 1.45}$
	ER+TRADES+Ours	$\mathbf{10.95_{\pm 0.8}}$	$\mathbf{51.63_{\pm 1.11}}$	$\mathbf{3.01_{\pm 0.24}}$	$\mathbf{2.79_{\pm 0.25}}$	$\mathbf{21.74_{\pm 0.6}}$	$\mathbf{44.81_{\pm 0.13}}$	$\mathbf{15.20_{\pm 0.08}}$	$\mathbf{12.92_{\pm 0.16}}$	$\mathbf{16.21_{\pm 0.06}}$	$\mathbf{12.33_{\pm 0.42}}$

Table S9. Adaptive Attack for ER+TRADES+ours. All the attack methods in the table incorporate the same logit calibration as in the training phase of our model. Forgetting is based on PGD-20. Results show our method still maintains decent robustness.

Dataset	Buffer Size	Class Incremental Setting			Task Incremental Setting		
		PGD-20	**AA**	**Forgetting**	**PGD-20**	**AA**	**Forgetting**
CIFAR-10	**200**	$23.99_{\pm 1.36}$	$41.76_{\pm 0.00}$	$26.43_{\pm 8.04}$	$58.18_{\pm 0.17}$	$58.32_{\pm 0.26}$	$15.40_{\pm 0.26}$
	5120	$27.50_{\pm 2.80}$	$35.29_{\pm 0.52}$	$1.53_{\pm 0.69}$	$65.59_{\pm 3.26}$	$60.69_{\pm 0.50}$	$5.89_{\pm 3.20}$
CIFAR-100	**500**	$9.91_{\pm 0.52}$	$20.61_{\pm 0.98}$	$13.10_{\pm 1.52}$	$26.84_{\pm 0.94}$	$28.20_{\pm 0.72}$	$17.13_{\pm 0.79}$
	2000	$13.68_{\pm 0.93}$	$25.85_{\pm 0.73}$	$15.33_{\pm 1.02}$	$34.83_{\pm 0.62}$	$35.72_{\pm 0.84}$	$11.59_{\pm 0.42}$

Table S10. GPU memory usage and training time per epoch on a single RTX 3090 GPU using Split-CIFAR10 with buffer size 200 and batch size 32. The results presented are the average values computed across tasks 2 through 5.

Methods	SGD+AT	Joint AT	ER+AT	DER+AT	DER+++AT	X-DER+AT	GSS+AT	ASER+AT	ER+TRADEs	ER+FAT
GPU memory/MB	2520	2524	2724	2807	2832	2904	2847	2723	2710	2808
Training time/s	350	1787	356	364	369	387	381	359	438	389

Methods	ER+LBGAT	ER+SCORE	ER+AT+Ours	ER+TRADES+Ours	GSS+AT+Ours	ASER+AT+Ours	X-DER+AT+Ours
GPU memory/MB	3722	2819	2758	2773	2853	2734	2984
Training time/s	508	447	358	439	396	360	389

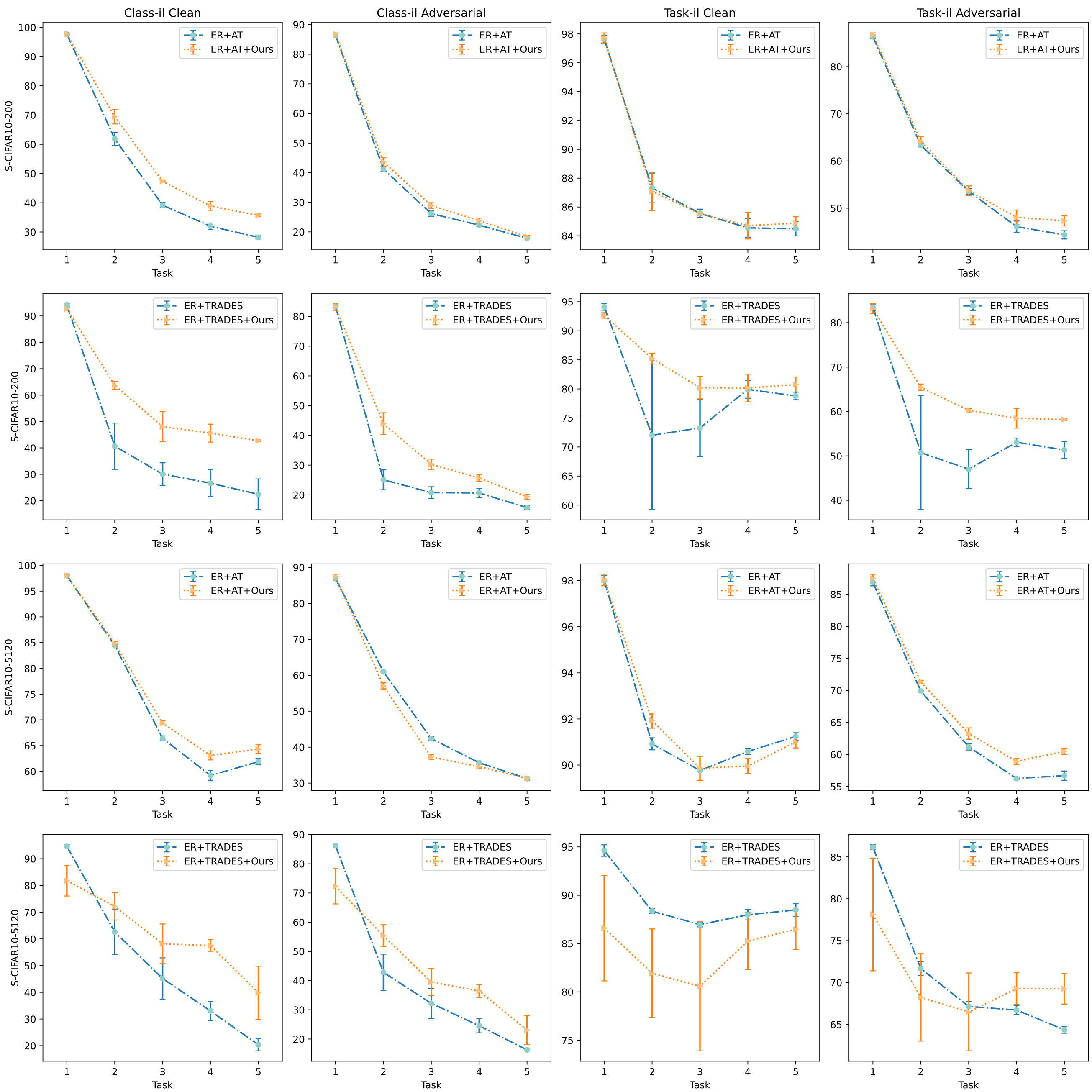

Figure S3. Accuracy curves on the Split-CIFAR10 with buffer size 200, and 5,120 settings. The plots' x-axis denotes the total number of tasks trained cumulatively up to each learning stage. The y-axis shows the average accuracy of the current task at each respective stage. The results demonstrate consistent improvements across most stages of continual learning when our proposed approach is combined with the baseline model.

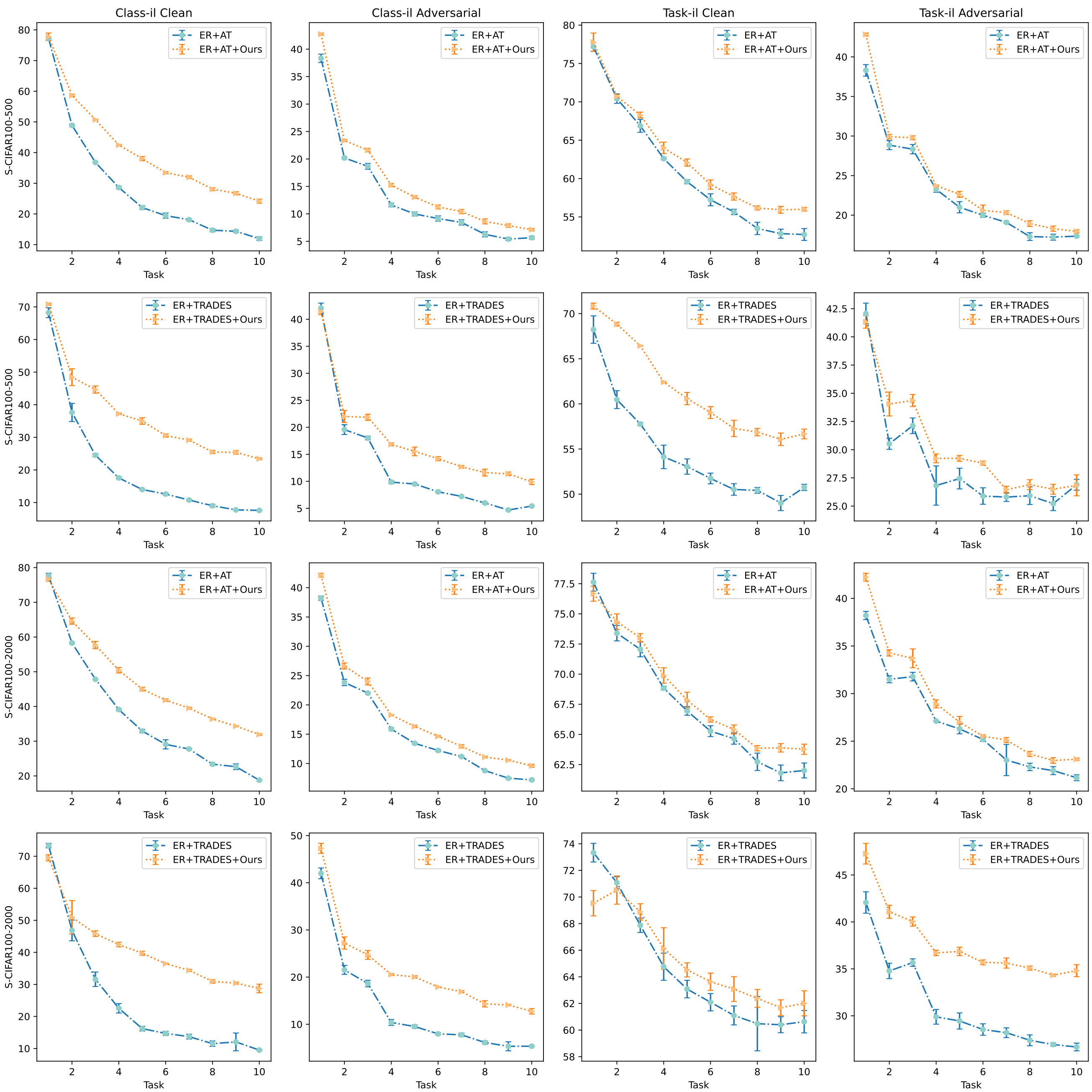

Figure S4. Accuracy curves on the Split-CIFAR100 with buffer size 500, and 2,000 settings. The plots' x-axis denotes the total number of tasks trained cumulatively up to each learning stage. The y-axis shows the average accuracy of the current task at each respective stage. The results demonstrate consistent improvements across most stages of continual learning when our proposed approach is combined with the baseline model.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 2

[2] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019. 3

[3] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. 2020. 5

[4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018. 5

[5] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 2

[6] Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended der-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 3

[7] 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. 1, 3, 8

[8] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. In *International Conference on Learning Representations*, 2022. 1

[9] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020. 2, 5

[10] Tianlong Chen, Sijia Liu, Shiyu Chang, Lisa Amini, and Zhangyang Wang. Queried unlabeled data improves and robustifies class-incremental learning. *Transactions on Machine Learning and Data Mining*, 2022. 1, 2

[11] Ting-Chun Chou, Jhih-Yuan Huang, and Wei-Po Lee. Continual learning with adversarial training to enhance robustness of image recognition models. In *2022 International Conference on Cyberworlds*, pages 236–242. IEEE, 2022. 2

[12] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2

[13] Jiequan Cui, Shu Liu, Liwei Wang, and Jiaya Jia. Learnable boundary guided adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15721–15730, 2021. 2, 3

[14] 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. 1

[15] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A Kai Qin, and Yuan He. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7506–7515, 2021. 2

[16] Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018. 1

[17] Yunhui Guo, Mingrui Liu, Yandong Li, Liqiang Wang, Tianbao Yang, and Tajana Rosing. Attacking lifelong learning models with gradient reversion. 2020. 1

[18] Ahmad Hassanpour, Majid Moradikia, Bian Yang, Ahmed Abdelhadi, Christoph Busch, and Julian Fierrez. Differential privacy preservation in robust continual learning. *IEEE Access*, 10:24273–24287, 2022. 1

[19] Hikmat Khan, Nidhal Carla Bouaynaya, and Ghulam Rasool. Adversarially robust continual learning. In *2022 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2022. 1

[20] Hikmat Khan, Pir Masoom Shah, Syed Farhan Alam Zaidi, et al. Susceptibility of continual learning against adversarial attacks. *arXiv preprint arXiv:2207.05225*, 2022. 1

[21] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 1

[22] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 2

[23] Lilly Kumari, Shengjie Wang, Tianyi Zhou, and Jeff Bilmes. Retrospective adversarial replay for continual learning. In *Advances in Neural Information Processing Systems*, 2022. 1

[24] Sungyoon Lee, Hoki Kim, and Jaewook Lee. Graddiv: Adversarial robustness of randomized neural networks via gradient diversity regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2645–2651, 2023. 2

[25] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 1

[26] Guoliang Lin, Hanlu Chu, and Hanjiang Lai. Towards better plasticity-stability trade-off in incremental learning: A simple linear connector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2022. 1

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2, 3

[28] Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation mod-

els. In *International Conference on Machine Learning*, pages 6640–6650. PMLR, 2020. 2

[29] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2023. 1

[30] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 1

[31] Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. *Advances in Neural Information Processing Systems*, 35:18599–18611, 2022. 4

[32] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. Boosting adversarial training with hypersphere embedding. In *Advances in Neural Information Processing Systems*, pages 7779–7792. Curran Associates, Inc., 2020. 5

[33] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021. 2, 3

[34] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. *International Conference on Machine Learning*, 2022. 2, 3, 7

[35] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. 1

[36] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990. 1

[37] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019. 1, 3

[38] Mohammad Rostami, Leonidas Spinoulas, Mohamed Hussein, Joe Mathai, and Wael Abd-Almageed. Detection and continual learning of novel face presentation attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14851–14860, 2021. 2

[39] 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. 1

[40] Rui Shao, Pramuditha Perera, Pong C Yuen, and Vishal M Patel. Open-set adversarial defense. In *European Conference on Computer Vision*, pages 682–698. Springer, 2020. 2

[41] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9630–9638, 2021. 3

[42] Stanford. Tiny imagenet challenge (cs231n). `http://tiny-imagenet.herokuapp.com/`, 2015. Accessed: Feb 21, 2023. 2

[43] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[44] Muhammad Umer, Glenn Dawson, and Robi Polikar. Targeted forgetting and false memory formation in continual learners through adversarial backdoor attacks. In *2020 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2020. 1

[45] Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Tiehang Duan, and Mingchen Gao. Improving task-free continual learning by distributionally robust memory evolution. In *International Conference on Machine Learning*, pages 22985–22998. PMLR, 2022. 1

[46] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022. 1

[47] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, 2023. 2

[48] Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, and Gholamreza Haffari. Pretrained language model in continual learning: A comparative study. In *International Conference on Learning Representations*, 2021. 1

[49] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8659–8668, 2021. 2

[50] Ju Xu, Jin Ma, Xuesong Gao, and Zhanxing Zhu. Adaptive progressive continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6715–6728, 2022. 1

[51] Zonghan Yang, Tianyu Pang, and Yang Liu. A closer look at the adversarial robustness of deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2022. 5

[52] Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang. Scalable and order-robust continual learning with additive parameter decomposition. *arXiv preprint arXiv:1902.09432*, 2019. 1

[53] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019. 2, 3

[54] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning*, pages 11278–11287. PMLR, 2020. 3

[55] Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15115–15125, 2022. 2

[56] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. *arXiv preprint arXiv:2302.03648*, 2023. 1

[57] Kaijie Zhu, Jindong Wang, Xixu Hu, Xing Xie, and Ge Yang. Improving generalization of adversarial training via robust critical fine-tuning. *arXiv preprint arXiv:2308.02533*, 2023. 2