Defense without Forgetting: Continual Adversarial Defense with Anisotropic & Isotropic Pseudo Replay

Yuhang Zhou[1], Zhongyun Hua[1,2*]

[1]Harbin Institute of Technology, Shenzhen,
[2]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies

{23B95105@stu, huazhongyun@}.hit.edu.cn

Abstract

Deep neural networks have demonstrated susceptibility to adversarial attacks. Adversarial defense techniques often focus on one-shot setting to maintain robustness against attack. However, new attacks can emerge in sequences in real-world deployment scenarios. As a result, it is crucial for a defense model to constantly adapt to new attacks, but the adaptation process can lead to catastrophic forgetting of previously defended against attacks. In this paper, we discuss for the first time the concept of continual adversarial defense under a sequence of attacks, and propose a lifelong defense baseline called Anisotropic & Isotropic Replay (AIR), which offers three advantages: (1) Isotropic replay ensures model consistency in the neighborhood distribution of new data, indirectly aligning the output preference between old and new tasks. (2) Anisotropic replay enables the model to learn a compromise data manifold with fresh mixed semantics for further replay constraints and potential future attacks. (3) A straightforward regularizer mitigates the 'plasticity-stability' trade-off by aligning model output between new and old tasks. Experiment results demonstrate that AIR can approximate or even exceed the empirical performance upper bounds achieved by Joint Training.

1. Introduction

Plenty of studies have demonstrated that high-performance deep neural networks (DNNs) are vulnerable to adversarial attacks [27, 31, 44], indicating that the addition of carefully designed but human-imperceptible perturbations to the inputs of a DNN can easily deceive the network and lead to incorrect predictions, a phenomenon known as adversarial attack. The existence of such attacks presents a significant threat to the deployment of the DNN-based systems. For instance, the elaborate physical adversarial attacks [3, 6, 23]

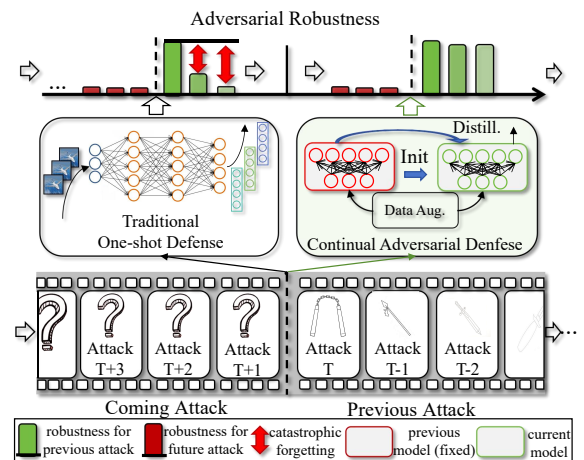
*Zhongyun Hua is the corresponding author.

Figure 1. The difference between one-shot defense and continual defense. The model diagram on the left presents the one-shot defense studies an isolated Min-Max process and implicitly assumes the potential attack is static. For a continual attack sequence, the indispensable adaptation process introduces additional challenge of catastrophic forgetting of previous attacks. Therefore, a deployable adversarial defense should be a life-one learning task rather than a one-shot task. We propose a self-distillation pseudo-replay baseline to alleviate the catastrophic forgetting against attack sequence, indicated by the model diagram on the right.

may cause the DNNs-based auto-drive systems to make incorrect judgments, potentially resulting in traffic accidents.

To ensure the reliability of DNNs in various scenarios, many defense methods have been proposed to maintain the robustness of DNN against adversarial attacks [16, 25, 34, 48, 54]. Existing defenses are always limited to the one-shot assumption, which indicates that the model enters a static process after a single defense training stage. However, in real deployment scenarios, new attacks occurs continuously [11, 13, 16, 22, 27] and even become a task sequence. As a result, defense model needs to constantly adapt to new attacks for adaptive robustness and turns into a life-long
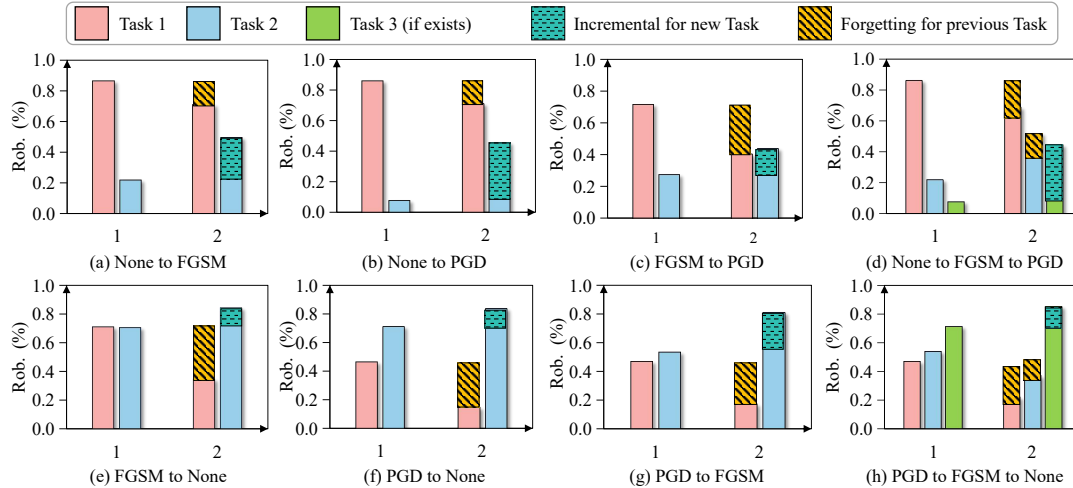
Figure 2. Catastrophic forgetting verification of one-shot defense model in continual defense scenario. The horizontal axis can be considered as a timestamp, where time '1' represents the model adapting to *TASK 1*, and time '2' represents the sequential adaptation to all attack tasks in the sequence. *TASK 1* and *TASK 2* depend on the specific sequence. For example, for the sub-first figure, *TASK 1* and *TASK 2* refer to None Attack and FGSM Attack, respectively.

learning [12, 29] rather than a one-shot task. Considering that DNNs are easily suffer from catastrophic forgetting [12, 29, 30], adapting to new attacks is unavoidable to result in suboptimal forgetting of previous one, which poses a new threat to the reliability of DNNs. Figure 1 illustrates the difference between one-shot and continual defense.

To demonstrate this challenge, we first explore the continual challenge of DNNs' adversarial robustness under continual attack sequences. We validate the catastrophic forgetting of standard adversarial training [27] on the attack sequences consisting of two and three attacks, as shown in Figure 2. The experiment is conducted on the CIFAR10 dataset with the backbone being Wide-ResNet-34 [51]. The validation experiment includes two settings: 'from easy to difficult' (i.e., FGSM to PGD) and 'from difficult to easy' (i.e., PGD to FGSM). For more validation experiments under different attack principles, varying attack intensities, and transferring between black and white box attacks), please refer to Section 1 of the supplementary materials. Obviously, adversarial training suffers from significant catastrophic forgetting under all attack sequences, and the forgetting becomes more severe under the 'difficult to easy' attack sequence. These results confirm our concern about DNNs' forgetting robustness to previous attacks as they constantly adapt to new ones, highlighting the need for defense with continual adversarial robustness.

To achieve robustness against continual adversarial, a potential solution is to employ Joint Training with all sequential attacks or fine-tune a pre-trained robust model with new attacks. However, the Joint Training strategy have challenges as the previous attack data of pre-trained models may

not access due to privacy protection and legal restrictions, making Joint Training difficult. Additionally, since new attacks are constantly emerging, the training cost linearly increases as the increase of attack sequence, making it inefficient to train all the data every time. For the fine-tuning strategy, poor plasticity can lead to forgetting or insufficient learning. Therefore, defense in the source-free continual paradigm [12, 29] is necessary under the scenario of data privacy, adversarial robustness and life-long learning.

In this paper, we propose *Anisotropic & Isotropic Replay (AIR)* as a baseline for continual adversarial defense. AIR combines isotropy and anisotropy data augmentation to alleviate catastrophic forgetting in continual adversarial defense scenarios within a self-distillation pseudo replay paradigm. The isotropic stochastic augmentation is beneficial for breaking specific adversarial patterns, and an isolated learning objectives is obtained for pseudo-replay data to prevent pattern collapse and self-contradiction, implicitly constraining the consistency between the new model and the previous model in the neighbor distribution of new attacks. This alignment indirectly aligns the model's output of new and previous attacks, considering that all adversarial data can be regarded as neighborhood samples of the raw data. The anisotropic mix-up augmentation provides the model with richer fusion semantics and connects the manifolds of previous and new attacks, originally spaced more apart. To further optimize the trade-off between plasticity and stability, we introduce an intuitive regularizer to optimize the model's generalization by constraining the hidden-layer feature of new attacks and pseudo-replay attacks to be mapped to the same feature cluster.

Our contribution can be summarized as:

• We first discuss, validate, and analyze the catastrophic forgetting challenge of adversarial robustness under continual attack sequences threat.

• We tackle the twin challenge of adversarial robustness and continual learnability by proposing AIR as an efficient self-training reply baseline. Through the pseudo self-distillation, parameters with similar activation for new and previous attacks are found and retained. AIR can also achieve an implicit chain consistency.

• We evaluate the performance of several classic and plug-and-play continual learning methods for continual adversarial attacks by combining them with adversarial training. Qualitative and quantitative experiments verify the feasibility and superiority of our AIR.

2. Related Work

2.1. Adversarial Attack & Defense

Adversarial Attack. Attacks on DNN models can be catrgorized into white box and black box attacks based on whether the attacker can access the target model or not. In the white box attack, the adversary can fully query and access various aspects of the target model, such as the model parameters, structure, and gradients. Mainstream white box attacks include gradient-based techniques (e.g., FGSM [16], BIM [22], MIM [13], PGD [27], and AA [11]), conditional optimization-based techniques (e.g., CW [8], OnePixel [43]), and classifier perturbations-based techniques (e.g., DeepFool [31]). On the other hand, black box attack occurs when the attacker has limited knowledge about the target model. This category can be further divided into score-based, decision-based, and transfer-based attacks. In score-based black box attack, the attacker can access to the probabilities (e.g., zoo [10]). Decision-based black box attacks operate under the constraint that the attacker can only obtain the one-hot prediction (e.g., Boundary attacks [5]). Transfer-based attack [13, 14, 26] typically involve crafting adversarial attacks using a substitute model, commonly employed to evaluate the adversarial robustness of DNNs.

Adversarial Defense. To maintain adversarial robustness under attacks, early adversarial defense are often heuristic, including input transformation [48], model ensemble [4], adversarial denoiser [40]. However, most of these models have been proven to benefit from unreliable obfuscated gradients [2]. Recently, adversarial training (AT) [16, 18, 27] and defensive distillation [15, 46, 46, 55] have become mainstream defense due to their essential robustness, and the latest research primarily focus on exploring their potential. In adversarial training, Jia *et al.* [18] proposed a learnable strategy, Mustafa *et al.* [32] enhanced adversarial robustness by perturbing feature representations, and

Pang *et al.* [33] fully utilized tricks to maximize the potential of AT. For defensive distillation, Wang *et al.* [45] introduced a bidirectional metric learning framework guided by an attention mechanism, and Zi *et al.* [56] proposed to fully exploit soft labels generated by a robust teacher model. The aforementioned defenses are limited to the one-shot settings and cannot adapt to new attack, resulting in insufficient robustness for the potential attack sequences. Some latest attempts to address this adaptive challenge include Test Time Adaptation Defense (TTAD) [41, 50], which considers continual adaptation to new attacks. However, TTAD only focuses on adaptation to unlabeled attacks on test data based on the current model, and ignores to alleviate the catastrophic forgetting of previous attacks. In this study, we propose a novel continual defense task where the defender considers both new and previous attacks.

2.2. Continual & Incremental Learning

DNNs tend to suffer from catastrophic forgetting, where adaptation to new tasks leads to a drop of performance for previous tasks. Continual learning [29], also known as incremental learning [12] and life-long learning [35], is considered a potential solution to atastrophic forgetting. In continual learning, tasks are roughly divided into 'Incremental task learning', 'Incremental class learning', and 'Incremental domain learning' based on the differences between new and previous tasks. For continual learning models, methods can be roughly categorized into replay-based [9, 38], parameter-isolation-based [1, 49], and regularization-based [21, 24, 36] methods, depending on how the task-specific information is stored and used throughout the sequential learning process. Replay-based methods can be further be divided into explicit replay [9] and pseudo replay [37]. Regularization-based methods can further be divided into data-focused [24] and prior-focused methods [21]. Parameter-isolation-based methods can be further divided into module allocation [1]) and additional modules [20]. Despite several attempts on basic tasks, continuous learning scenarios have not yet been deeply explored in adversarial defense. With the increasing variety of attacks, reliable DNN models necessitate adaptable lifelong defense against attack sequences.

3. Methodology

3.1. Overview

Continual adversarial defense resembles 'incremental domain learning' more than 'incremental task learning' or 'incremental class learning' [29]. On one hand, the output space of the attack sequence is fixed for one dataset, allowing the defense model for each attack to share a common classifier. On the other hand, even though the distribution at the feature level may shift, the low-level semantics of
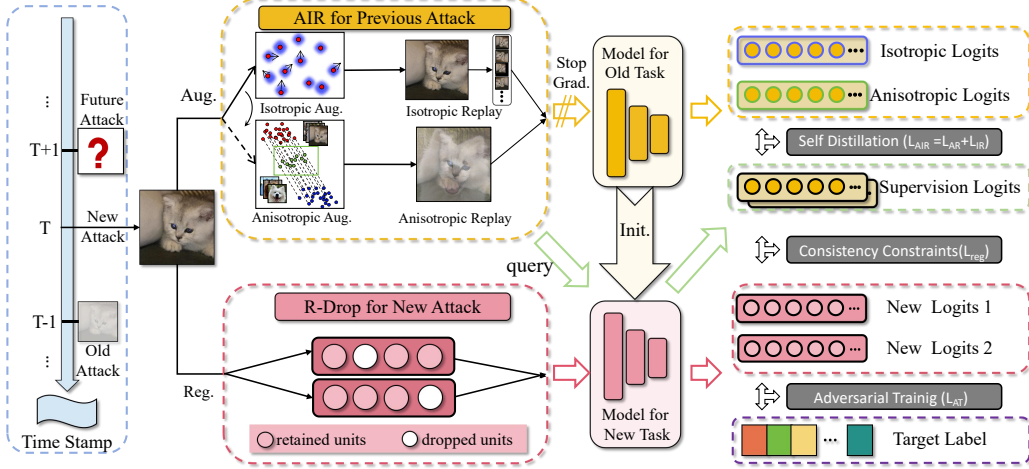
Figure 3. Framework of our AIR. The upper module (in yellow block) consists of the anisotropic replay module and isotropic replay module, aiming to maintain the memory of old tasks. The lower module (in red block) is the vanilla adversarial training with R-Drop for new attacks. The three main loss functions are highlighted in the gray circular box.

adversarial samples remain invariant, which means that the adversarial samples of cats are always visually the same as cats (constrained by the definition of adversarial samples) and can be considered as the neighbor of the raw data. As a result, complex prior knowledge and large memory may not be necessary to obtain replay data.

However, directly utilizing new attack data as pseudo replay data may be insufficient. On the one hand, new data needs to be mapped to real labels. Simultaneously aligning the output of the new data with the output of the teacher model and the real label may lead the model to be trapped in a self-contradiction dilemma. On the other hand, the new data are still adversarial, making it challenging to for the defense model to fit them without forgetting. Hence, we propose a composite data augmentation scheme to establish an efficient self-distillation pseudo replay paradigm.

3.2. Task Definition.

We first provide a definition of continual adversarial defense. In continual adversarial defense, the model learns an attack sequence $\mathcal{A} = \{A_1, ...A_t, ...A_N\}$ one by one, and the attack ID is available during both training and testing stages. We assume that each attack A_t has a manually labeled training set and test set, denoted as $A_t^{train} = \{(x_t^i, y_t^i), i = 1, ..., n_t^{train}\}$, where y_t^i is the real label, and n_t^{train} is the number of the training data. The test set can be defined as A_t^{test} in the similar way. A reliable defense model should learn a new attack A_t without forgetting the previous attacks $\{A_1, ...A_{t-1}\}$.

3.3. Self-distillation Pseudo Replay

By aligning the outputs of the current and previous models on pseudo replay data, the current model can learn the map-

ping preferences of the previous model, indirectly maintaining the mapping relationship between the previous data-label pair. Generally, assuming we have a model f_{w_t} parameterized by w_t at time t, the self distillation pseudo replay can be formalized as:

$$\mathcal{L} = \mathcal{L}_{vanilla} + \mathcal{L}_{dis}, \tag{1}$$

where $\mathcal{L}_{vanilla}$ is the classification loss of the current model for new data, commonly known as the cross entropy loss. In continual adversarial defense, $\mathcal{L}_{vanilla}$ refers to the vanilla adversarial training loss for a new attack [27] and we represent it as $\mathcal{L}_{AT}$ in the following text.

The $\mathcal{L}_{dis}$ represents the self-distillation loss from the previous model to the current model on pseudo replay data, and can be formalized as:

$$\mathcal{L}_{dis} = D(f_{w_{t-1}}(X_t'), f_{w_t}(X_t')), \tag{2}$$

where D is the diversity measure, commonly known as the KL Divergence, $f_{w_{t-1}}$ is the optimal previous model at time $t - 1$, and X_t' is the pseudo data at time t.

3.4. Isotropic Pseudo Replay

To align the outputs of the new and old models, we create independent data for pseudo replay based on the current adversarial samples. Assume that we already have a certain batch of data X_t, the neighborhood samples of new data can be obtained by:

$$X_t^{IR} = \mathcal{T}(X_t + \lambda \cdot r), \tag{3}$$

where λ is a hyper-parameter, r is a stochastic perturbation sampled from the Gaussian distribution, and $\mathcal{T}$ is a random

augmentation operator that includes random rotation, cropping, flipping, and erasing. In this 'perturbation on attack' way, specific distributions (e.g., pixel or texture [17, 28]) in the adversarial perturbation are somehow broken, wakening the attack ability. Simultaneously, independent pseudo replay data unrealted to the current attack is obtained. The augmented pseudo replay data does not deviate from the raw semantics, and its potential label is also fixed. Hence it is called isotropic replay (IR), and the IR loss can be formalized as:

$$\mathcal{L}_{IR} = KL\_Div(f_{w_t}(X_t^{IR}), f_{w_{t-1}}(X_t^{IR})). \quad (4)$$

3.5. Anisotropic Pseudo Replay

To further obtain a compact and uniform manifold of the replay data and ensure a nontrivial solution, we introduce an anisotropic data augmentation scheme: $mix-distill$, which evolves from $mixup$ [52]. We randomly shuffle the current batch data X_t to obtain a new batch of data $X_t^{shuffle}$:

$$X_t^{AR} = \alpha \cdot X_t + (1-\alpha) \cdot x_t^{shuffle}, \quad (5)$$

where α is a stochastic mixing weight sampled from $U[0.3, 0.7]$. The label for the replay is also required. The pseudo label for the mixed data is obtained in the same way as in the vanilla $mixup$. However, real labels are agnostic in pseudo replay framework and we attempt to obtain the mixed labels in two ways:

Mixed query labels. Referring to the standard $mixup$, we collect the output logits of the teacher model (previous model) for the two components of the mixed data, and mix them with the same weight as the data:

$$y_{dis} = \alpha \cdot f_{w_{t-1}}(X_t) + (1-\alpha) \cdot f_{w_{t-1}}(X_t^{shuffle}). \quad (6)$$

However, the label mixing strategy in the standard $mixup$ sometimes leads to training collapse. This is because pseudo labels are inherently inaccurate, and mixing suboptimal pseudo labels leads to error accumulation, making it difficult to align the preference of teachers and students. The nonlinearity of the model may also amplify the mapping shift of the pseudo-replay self-distillation model.
Query label of the mixed data. Based on the above analysis, we directly align the output preferences of the teacher-student model for mixed samples, and the AR losses can be expressed as

$$\mathcal{L}_{AR} = KL\_Div(f_{w_{t-1}}(X_{AR}), f_{w_t}(X_{AR})). \quad (7)$$

The nature of 'anisotropy' is reflected not only in the feature mixing provided by pixel-level interpolation (such as the mixing of $lion$ and $tiger$, which may introduce new semantics like $liger$), but also in the indirect combination of supervise labels. In this way, the intra-class gap is initially
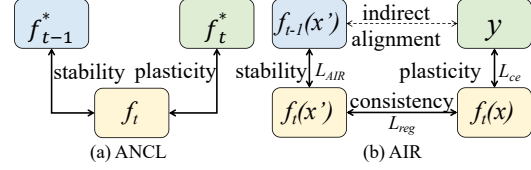

Figure 4. Achievement of chain consistency of our AIR in end-to-end paradigm. Models with $*$ superscripts (such as f_t^*) are additionally trained independently of the main pipeline.

filled, and the data manifold becomes uniform. This also provides the model with the ability to generalize to unfamiliar and unseen semantics, as such semantics may arise in future tasks. More specifically, adversarial samples may have already introduced new semantics to the original data, and the mixing between adversarial samples further explores a richer internal maximization process.

3.6. Regularization for the Trade-off

A common dilemma in continual learning is the trade-off between stability and plasticity. For the 'domain-incremental-like' continuous defense, a natural shortcut to optimizing the trade-off is to assign all attacks under a category to the same cluster. This approach elegantly optimizes the trade-off and achieves consistent optimization of attack sequences. In our AIR, augmentation data is considered as replay data to query the previous model and can also be considered as a neighborhood sample of new data. We propose aligning the outputs of the two as follows:

$$\mathcal{L}_{reg} = \frac{1}{2}(KL(f_{w_t}^1(x_t)||f_{w_t}^1(x_t')) \\ + KL(f_{w_t}^2(x_t)||f_{w_t}^2(x_t'))), \quad (8)$$

where x_t' is the isotropic augmentation replay data. The alignment process is implemented in the R-Drop [47] way, where $f_{w_t}^1$ and $f_{w_t}^2$ represent that the inputs are fed into the model with random Dropout [42] twice, respectively. On the one hand, a model with R-Drop will learn consistent outputs from different local features without overfitting on specific features. On the other hand, this constraint achieves an indirect chain alignment. Compared to the regularization from middle to the both sides in ANCL [20], it resembles an alignment from both sides to the middle. Besides, AIR does not require additional auxiliary networks, and the alignment can be implemented in an end-to-end process (see in Figure 4). The R-Drop is also applied to the AT for new attack with a minor probability.

3.7. Final Model

Integrally, the overall loss of AIR can be formalized as:

$$\mathcal{L}_{AIR} = \mathcal{L}_{AT} + \lambda_{SD} \cdot (\mathcal{L}_{IR} + \mathcal{L}_{AR}) + \lambda_{Reg} \cdot \mathcal{L}_{Reg}, \quad (9)$$

		Transfer between two attacks										
Datasets	Tasks	None to FGSM		FGSM to None		None to PGD		PGD to None		FGSM to PGD		PGD to FGSM	
		Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
MNIST	Vanilla AT [27]	95.18	98.55	83.97	98.86	94.22	90.01	3.72	98.59	96.48	94.71	2.56	96.96
	EWC [21]	98.83	96.63	98.18	97.85	97.35	87.32	91.97	98.85	95.26	**95.86**	94.77	96.90
	Feat. Extraction [24]	98.16	89.23	97.46	98.80	12.72	11.35	95.23	98.80	96.94	73.61	95.23	97.93
	LFL [19]	98.85	97.02	90.54	98.80	97.32	87.52	33.84	98.71	95.84	91.87	25.05	**98.40**
	AIR (ours)	**99.37**	**98.84**	98.18	**98.84**	**98.89**	**94.26**	**95.93**	**99.06**	**97.45**	95.67	**96.25**	97.93
	Joint Training [24]	99.11	98.52	98.52	99.11	99.35	95.44	95.44	99.35	96.72	94.29	94.29	96.72
CIFAR10	Vanilla AT [27]	70.60	49.30	34.90	83.83	71.09	45.52	15.19	83.59	34.90	35.21	17.14	60.24
	EWC [21]	72.66	49.17	**43.85**	82.62	69.38	41.46	30.25	61.70	48.63	40.53	24.44	45.18
	Feat. Extraction [24]	67.69	35.11	45.27	82.13	40.04	30.90	**45.54**	75.02	52.85	24.88	42.51	44.54
	LFL [19]	74.23	50.17	42.77	78.59	67.31	42.76	28.27	**80.59**	51.98	43.30	24.18	46.71
	AIR (ours)	**76.73**	**51.48**	42.32	**82.85**	**75.53**	**45.14**	41.21	77.02	**53.39**	**44.12**	**43.00**	**52.26**
	Joint Training [24]	86.10	47.65	57.65	86.10	72.58	44.86	44.86	72.58	49.81	42.56	42.56	49.81
CIFAR100	Vanilla	42.27	20.67	25.98	50.26	40.58	17.31	20.21	47.47	24.08	19.03	20.89	30.47
	EWC [21]	50.04	22.43	**29.13**	45.12	**48.45**	16.61	19.21	44.66	22.98	18.00	20.16	24.32
	Feat. Extraction [24]	37.02	8.35	23.62	47.68	11.46	4.96	20.70	41.42	23.63	18.22	19.54	24.08
	LFL [19]	28.61	15.30	37.48	49.06	19.19	13.36	20.08	43.62	25.49	15.77	19.19	23.85
	AIR (ours)	**50.77**	**24.32**	27.47	**50.67**	47.88	**21.41**	**22.05**	**45.61**	**27.59**	**23.19**	**23.40**	**27.51**
	Joint Training [24]	56.44	35.88	35.88	56.44	46.01	22.54	22.54	46.01	35.27	21.45	21.45	35.27

Table 1. Adaptation between two attacks for different defense methods. Each method is combined with vanilla adversarial training and the Joint Training is considered to be the empirical upper bound marked with a gray background. The best performance, excluding Joint Training, is highlighted in bold.

where λ_{SD} and λ_{Reg} represent the hyper-parameters for self-distillation and regularization respectively. The AIR model will adapt to new attack while aligning with the previous model on the augmentation neighbor data, and the overall structure of AIR can be seen in Figure 3.

4. Experiments

4.1. Experimental Setup

Datasets and Backbones. Three commonly used datasets for adversarial attack & defense are explored:

• MNIST: Following the settings in [7, 53], we employ the smallCNN architecture, consisting of four convolutional layers and three fully-connected layers. Typically, we set the perturbation parameter $\epsilon = 0.3$, perturbation step size $\eta_1 = 0.01$, number of iterations $K = 40$, learning rate $\eta_2 = 0.01$, and batch size $m = 128$.

• CIFAR10 / CIFAR100: Following the settings in [27, 53], we employ the WRN-34-10 and WRN-34-20 architecture [51] for CIFAR10 and CIFAR100 respectively. We set the perturbation parameter $\epsilon = 8/255$, perturbation step size $\eta_1 = 2/255$, number of iterations $K = 10$, learning rate $\eta_2 = 0.1$, and batch size $m = 128$.

Evaluation Protocol. We set attack sequences with lengths of 2 and 3, following the setting of existing continual learning works [19, 24, 36], including 'from hard to easy attack' and 'from easy to hard attack' strategies. The classic FGSM attack [16] is selected as the easy attack, while the PGD attack [27] is chosen as the hard attack. Considering that all attacks originate from benign samples, these benign samples are also considered as a base task. The adversarial training corresponding to these above two types of attacks serves as the adaptation processes to specific attacks. We do not choose the advanced AutoAttack [11], as it is commonly used for evaluation rather than defensive training, and there is no specifically designed adaptation method for AutoAttack to our best knowledge yet. Additionally, we will explore attack sequences with different budgets of PGD.

4.2. Experimental Results

Attack Sequences of Different Lengths. The continual defense results of the attack sequences with a length of 2 are shown in Table 1, while that of the attack sequences with a length of 3 are shown in Table 2 (for MNIST), Table 3 (for CIFAR10) and Table 4 (for CIFAR100). Regardless of the length or arrangement of the attack sequence, the vanilla model consistently suffers from catastrophic forgetting. The internal maximization pattern and the 'ridges' reached by different attacks may vary, leading to a distribution shift in model parameters. Additionally, the 'from difficult to easy' attack sequence appears to exhibit more catastrophic forgetting compared to the 'from easy to difficult' sequence. This discrepancy may be attributed to the adaptation process to simple attacks being more trivial, making the model more prone to overfitting and trapping in local optima. According to [39], the essence of adversarial training is smoothness regularization, and our AIR may have added stronger smoothness constraints compared on the ba-

MNIST: None & FGSM & PGD						
Tasks	None to FGSM to PGD			PGD to FGSM to None		
	Task1	Task2	Task3	Task1	Task2	Task3
Vanilla AT [27]	95.97	97.55	96.90	2.66	71.49	98.73
EWC [21]	98.97	96.47	92.91	89.67	95.43	99.11
Feat. Extra [24]	11.71	11.35	11.38	90.92	95.99	99.18
LFL [19]	99.36	94.80	88.15	10.06	89.51	98.97
AIR (ours)	**99.39**	**97.21**	**94.54**	91.55	**97.34**	**99.33**
Joint Training [24]	99.11	97.07	94.76	**94.76**	97.07	99.11

Table 2. Results among None, FGSM, and PGD on MNIST.

CIFAR10: None & FGSM & PGD						
Tasks	None to FGSM to PGD			PGD to FGSM to None		
	Task1	Task2	Task3	Task1	Task2	Task3
Vanilla AT [27]	70.93	40.70	44.04	21.87	36.80	84.59
EWC [21]	68.31	45.09	36.57	27.32	51.66	75.22
Feat. Extraction [24]	37.95	53.28	40.62	**44.42**	**53.63**	73.77
LFL [19]	74.21	52.42	42.89	22.12	46.27	76.37
AIR (ours)	**75.75**	**53.51**	**43.12**	42.35	52.44	**76.66**
Joint Training [24]	70.62	51.35	44.36	44.36	51.35	70.62

Table 3. Results among None, FGSM, and PGD on CIFAR10.

CIFAR100: None & FGSM & PGD						
Tasks	None to FGSM to PGD			PGD to FGSM to None		
	Task1	Task2	Task3	Task1	Task2	Task3
Vanilla	42.01	22.30	17.80	19.50	22.62	47.04
EWC [21]	**48.35**	22.54	16.59	19.88	24.48	44.83
Feat. Extraction [24]	2.17	1.67	4.63	21.01	24.84	41.88
LFL [19]	29.97	10.39	8.96	19.94	24.36	**45.41**
AIR (ours)	47.08	**27.34**	**23.04**	**23.12**	**27.04**	44.16
Joint Training [24]	45.33	30.23	21.25	21.25	30.23	45.33

Table 4. Results among None, FGSM, and PGD on CIFAR100.

CIFAR10: Weak Attack & Strong Attack				
Tasks	Weak to Strong		Strong to Weak	
	Task1	Task2	Task1	Task2
Vanilla	36.59	39.35	10.89	42.56
AIR (ours)	**45.70**	**37.20**	**29.09**	**42.87**
Joint Training [24]	37.03	36.57	36.57	37.03

Table 5. Results between Weak & Strong PGD on CIFAR10.

sic traditional one-shot adversarial training to prevent output shifts in the case of continual input shift. Another issue is that the performance of our AIR may fluctuate. One can select a better model by monitoring the training online. A more detailed analysis of the results can be found in Section 2 of the supplementary materials.

Transfer between different attack budgets. Attacks with different budgets present different threats, which may seem counterintuitive. People may generally expect models adversarially trained with stronger internal maximization to be robust to weaker attacks. However, our findings reveal that the model exhibits different preferences for internal maximization with different budgets. This means that sequences composed of attacks with different intensities can lead to catastrophic forgetting. We briefly explore this issue on CIFAR10, setting attack budgets for strong and weak attacks to 8/255 and 80/255, respectively. Table 5 shows our ex-

perimental results and our AIR successfully alleviates the catastrophic forgetting caused by attack sequences with different attack budgets.

Ablation study. We conducted ablation analysis of AIR on CIFAR10 with the 'hard to easy' attack sequence, which suffers from more serious forgetting and the comparisons are more significant. Figure 5 shows the ablation study results. Essentially, each of our proposed modules contributes improvements for previous attacks (PGD attack in this case). The AR and IR modules individually enhance the performance against the previous attack, while the composite regularizer provides an overall increment for both previous and new attacks. Due to the trade-off involving 'plasticity-stability' and 'robustness-precision', our AIR, like other continual learning methods, also experiences performance degradation for new tasks. However, this sacrifice is necessary and cost-effective, as it results in effective improvement on previous attacks, aligning with the 'small pain but great gain' philosophy. Moreover, the overall performance of the model for both new and previous tasks steadily increases. In summary, the ablation analysis demonstrates the effectiveness of our designs in AIR.

Discussion. Our AIR sometimes can even outperforms Joint Training (JT). JT is typically considered as the empirical upper bound of continual learning, while our AIR can approach or even surpass JT in a memory-free manner. This may be attributed to the fact that the old attacks can be considered as the pre-training for new attacks. However, this superiority appears to be conditional: pre-training from easy to difficult attacks exhibits a better regularization effect on subsequent tasks, indicating that features from simpler tasks may be more generalizable. Utilizing previous task knowledge to enhance the learning of subsequent tasks may be one of the potentials of continual defense.

4.3. Feature Distribution

Our AIR tends to homogenize different attacks of the same category. In one-shot vanilla training, the model often allocates different clusters for different attacks of the same class, as observed in Figure 6. This partially explains the forgetting mechanism in continual adversarial defense. While each class under the new attack is clustered, it does not share the same cluster as the previous attack. Consequently, attacks with the same label become isomerized.
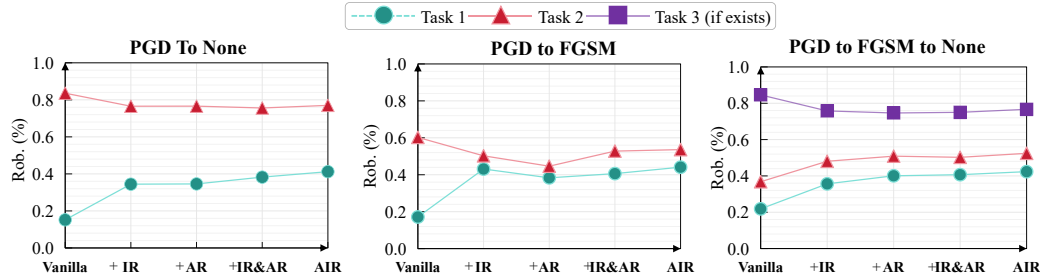
Figure 5. Ablation analysis of the 'from hard to easy' attacks on CIFAR10. We reported its results after learning the whole attack sequence.
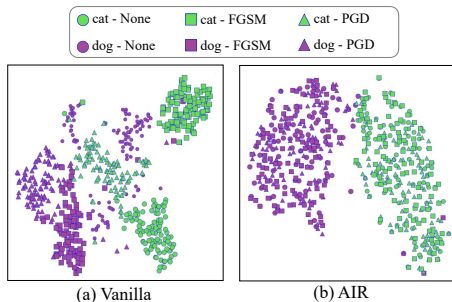

Figure 6. T-SNE diagram of features encoded by vanilla AT and AIR on CIFAR10. The proposed AIR is able to encode all attacks in the sequence of the same category into one shared cluster.

For instance, the features of FGSM and PGD attack of a certain label (e.g., 'dog') are assigned to different clusters, when ideally they should be in the same cluster. The adaptation learning of new clusters inevitably leads to the forgetting of old clusters. Intuitively, our AIR aligns the feature distribution of different attacks belonging to the same label into one cluster. This may benefit from the implicit alignment in the IR module and the chain regularizer, which aligns the output of new and previous models on the neighbor of the new data. This indirect alignment further harmonizes the feature distribution of new and previous models for different attacks of the same category, which provides an explanation beyond parameters regularization for AIR.

4.4. The Twinship Between Two Trade-offs

A common dilemma in adversarial defense is the trade-off between accuracy and robustness. Similarly, the 'plasticity-stability' dilemma in continual adversarial defense offers a novel perspective on the 'accuracy-robustness' trade-off. As the model transfers from benign data to adversarial data, the accuracy of benign tasks inevitably declines, reflecting the 'accuracy-robustness' dilemma. This reveals that the two dilemmas share similar insights: an excessive attention on the min-max process causes the model to forget relatively easy benign samples. Through the indirect alignment of the output preferences of old and new models, the pseudo-replay framework in our AIR can alleviate

the forgetting problem of benign samples without accessing the original data. This alignment can be interpreted as optimizing the trade-off between accuracy and robustness. Such unified perspective reveals that ideas to alleviate the 'accuracy-robustness' dilemma may also be effective in mitigating the forgetting problem in continual adversarial defense. In defense communities, a common way to address the 'accuracy-robustness' trade-off is TRADES [53], which aligns the output of clean and adversarial samples. However, querying previous data in continual defense is not feasible, making explicit alignment process challenging. Actually, our analysis indicates that AIR aligns the output preferences of new and old attacks in an indirect chain-like manner, which is similar to an implicit form of TRADES.

5. Conclusion and Outlook

In this paper, we first explore the challenge of achieving continual adversarial robustness under attack sequences, and verify that adaptation to new attacks can lead to catastrophic forgetting of previous attacks. Subsequently, we propose AIR as a memory-free continual adversarial defense baseline model. AIR aligns the outputs of new and old models in the neighborhood distribution of current samples and learns richer mixed semantic combinations to enhance adaptability to unknown semantics. An intuitive but efficient regularizer optimizes the generalization of multi-trade-offs in a chain-like manner.

One limitation of AIR is that it overlooks the regularization effect of previous knowledge for new tasks, where the previous tasks may act as pre-training. Additionally, we observed that in the 'from easy to difficult' attack sequence, AIR sometimes performs better in new tasks. There remains significant research space for continual adversarial defense.

6. Acknowledgments

References

[1] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3366–3375, 2017. 3

[2] 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. 3

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning*, pages 284–293. PMLR, 2018. 1

[4] Alexander Bagnall, Razvan Bunescu, and Gordon Stewart. Training ensembles to detect adversarial examples. *arXiv preprint arXiv:1712.04006*, 2017. 3

[5] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017. 3

[6] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 1

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 6

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 3

[9] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, P Dokania, P Torr, and M Ranzato. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019. 3

[10] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based blackbox attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017. 3

[11] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pages 2206–2216. PMLR, 2020. 1, 3, 6

[12] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2021. 2, 3

[13] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018. 1, 3

[14] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 3

[15] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3996–4003, 2020. 3

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 3, 6

[17] Binxiao Huang, Chaofan Tao, Rui Lin, and Ngai Wong. What do adversarially trained neural networks focus: A fourier domain-based study. *arXiv preprint arXiv:2203.08739*, 2022. 5

[18] 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. 3

[19] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. *arXiv preprint arXiv:1607.00122*, 2016. 6, 7

[20] Sanghwan Kim, Lorenzo Noci, Antonio Orvieto, and Thomas Hofmann. Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11930–11939, 2023. 3, 5

[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. 3, 6, 7

[22] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1, 3

[23] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018. 1

[24] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 3, 6, 7

[25] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018. 1

[26] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019. 3

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2, 3, 4, 6, 7

[28] Shishira R Maiya, Max Ehrlich, Vatsal Agarwal, Ser-Nam Lim, Tom Goldstein, and Abhinav Shrivastava. A frequency perspective of adversarial robustness. *arXiv preprint arXiv:2111.00861*, 2021. 5

[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, 2022. 2, 3

[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. 2

[31] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016. 1, 3

[32] Aamir Mustafa, Salman H Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Deeply supervised discriminative learning for adversarial defense. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9): 3154–3166, 2020. 3

[33] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020. 3

[34] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016. 1

[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. 3

[36] Amal Rannen, Rahaf Aljundi, Matthew B Blaschko, and Tinne Tuytelaars. Encoder based lifelong learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1320–1328, 2017. 3, 6

[37] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. 3

[38] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[39] Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 6

[40] Shiwei Shen, Guoqing Jin, Ke Gao, and Yongdong Zhang. Ape-gan: Adversarial perturbation elimination with gan. *arXiv preprint arXiv:1707.05474*, 2017. 3

[41] Changhao Shi, Chester Holtz, and Gal Mishne. Online adversarial purification based on self-supervision. *arXiv preprint arXiv:2101.09387*, 2021. 3

[42] HintonG SrivastavaN, Alex Krizhevsky, et al. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929, 2014. 5

[43] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. 3

[44] 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. 1

[45] Hong Wang, Yuefan Deng, Shinjae Yoo, Haibin Ling, and Yuewei Lin. Agkd-bml: Defense against adversarial attack by attention guided knowledge distillation and bi-directional metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2021. 3

[46] 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. 3

[47] Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905, 2021. 5

[48] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017. 1, 3

[49] Ju Xu and Zhanxing Zhu. Reinforced continual learning. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[50] Zhaoyuan Yang, Zhiwei Xu, Jing Zhang, Richard Hartley, and Peter Tu. Adaptive test-time defense with the manifold hypothesis. *arXiv preprint arXiv:2210.14404*, 2022. 3

[51] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 2, 6

[52] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[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. 6, 8

[54] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4480–4488, 2016. 1

[55] 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. 3

[56] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16443–16452, 2021. 3