Supplementary Material for “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

1. Additional Verification Experiments

As supplementary experiments to the catastrophic forgetting verification experiments introduced in the main text, we further investigate the performance of vanilla adversarial training against attack sequences in a diverse range of settings, as outlined below:

- PGD [6] & CW [1]. CW attack is a conditional-optimization based attack with different principles compared to PGD attack. We evaluate it, along with PGD, to verify forgetting in **attack sequence with different attack principles**. However, training CW attack online presents a significant challenge due to its substantial computational cost. Additionally, there is a lack of specifically designed adaptation methods for CW within the defense community [5]. Therefore, the defense model is adapted to CW attack after the attack data (including the training set and testing set) is pre-crafted.
- PGD [6] & AA [2]. AutoAttack (AA) is one of the most advanced attacks for evaluating the adversarial robustness of DNN model, incorporating both black box and white box attack strategies. We evaluate AA in a black box manner, along with PGD in a white box manner, to verify robustness forgetting against **the attack sequence consists of white box and black box attacks**. A substitute model (ResNet18 [3] in our experiments) is pre-trained to generate AA samples, following the commonly used black box attack approach [10].
- PGD [6] & DDN [8]. The decoupled direction and norm attack (DDN) concepts from CW and PGD attacks, and is essentially a white box attack. **As a supplement verification to the black box AA**, we also evaluated DDN in a black box manner along with PGD in a white box manner. A substitute model (ResNet18 [3]) is pre-trained to craft DDN samples.
- PGD [6] with **different attack budget**. Finally, it is a common but inaccurate perception that models defended with a larger budget attack should be robust to smaller budget attack. However, different budgets may introduce different inner maximization biases, leading to distinct ridges. Simply, models defended against a specific attack budget may not always exhibit optimal robustness against attacks with different budgets. We evaluated PGD with the attack budgets of 8/255 and 80/255 to investigate this challenge.

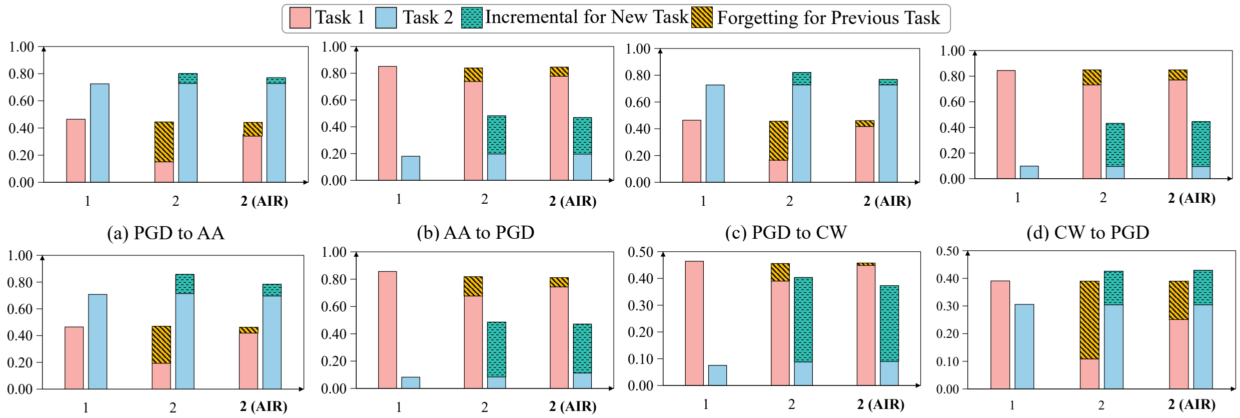
The verification results are presented in Figure 1, confirming that the defense model experiences catastrophic forgetting in all settings. However, the adaptive processes to CW, DDN and AA attacks are relatively straightforward, involving naive off-line adversarial training. In reality, there is a lack of specialized and rigorous adaptation methods in the community to these attacks. Consequently, for the evaluation of the continual defense model in the main text, we primarily focus on the FGSM attack and PGD attack with mature defense strategies (AT for FGSM and Madry's AT for PGD).

As a supplementary exploration in addition to the experimental section in the main text, we further investigated the performance of our AIR under these attack sequences in a naive offline adversarial training manner. The bar chart marked by subscript '**2 (AIR)**' illustrates the results. The AA, CW, and DDN attacks are considered relatively simple attacks in the experiments, as they are frozen after being crafted (either pre-crafted or obtained with a fixed substitute model). The model exhibits less forgetting for the 'easy to hard' attack sequences (e.g, CW to PGD) compared to the more severe forgetting observed for the 'hard to easy' attack sequences (e.g., PGD to CW). The proposed AIR consistently mitigates the model's forgetting of previous tasks (indicated by the yellow bar chart with diagonal stripes) and maintains improvements for the new tasks (shown in the blue bar chart with horizontal stripes), achieving a better 'stability-plasticity' trade-off.

2. Discussion on the Feature Extraction

Feature extraction is a common transfer-learning method where only the last fully connected layers are fine-tuned

[*] Zhongyun Hua is the corresponding author.



(a) PGD to AA (b) AA to PGD (c) PGD to CW (d) CW to PGD

(e) PGD to DDN (f) DDN to PGD (g) Weak PGD to Strong PGD (h) Strong PGD to Weak PGD

Figure 1. Supplementary catastrophic forgetting verification in a one-shot defense model within a continual defense scenario. The horizontal axis is a timestamp, where time '1' represents the model's adaptation to *TASK 1*, and time '2' represents the sequential adaptation to all attack tasks in the given sequence. The '**2 (AIR)**' represents the performance of our AIR after adapting to the entire sequence. *TASK 1* and *TASK 2* depend on the particular sequence. For example, in the first sub-figure, *TASK 1* and *TASK 2* refer to PGD attack and AA attack, respectively.

for adapting to new tasks, while other modules, such as convolutional layers, remain fixed as the feature extractor. This technique has been extensively evaluated in transfer learning![7] and continual learning [4]. However, in continual adversarial defense, feature extraction shows intriguing insights.

Poor performance against 'easy to hard' attack sequences. As shown in Tables 1-4 in the main text, the limited plasticity of the feature extraction model hinders its ability to learn more challenge attacks, such as 'none to PGD' in CIFAR100.

Excellent performance against 'hard to easy' attack sequences. However, feature extraction model performs well against 'hard to easy' attack sequences, even though only the FC layers are fine-tuned. The above results suggests that:

(a) Adversarial attacks tend to perturb the entire model, including the shallow representation, rather than targeting the classifier alone. This means that, for models without defense or those only defended with weak attacks, fine-tuning the classier alone is insufficient to achieve robustness, as the features encoded by the weak robust extractor are unreliable.

(b) Complex robust representations exhibit a degree of backward compatible with weak robust representations. However, additional fine-tuning is necessary for optimal performance. This observation is further supported by the results of the 'strong PGD & weak PGD' verification experiments, where the forgetting in 'weak PGD to strong PGD' sequence is not as severe as in the 'strong PGD to weak PGD' sequence. However, to achieve enhanced robustness against the weak attacks, an additional adaptation process is indispensable.

PGD & FGSM & Patch & AA & None										
Tasks	from PGD to None					from None to PGD				
	Task1	Task2	Task3	Task4	Task5	Task1	Task2	Task3T	Task4	Task5
Vanilla	24.62	43.16	34.23	67.97	**75.20**	**70.60**	**71.68**	40.04	36.33	40.04
AIR(ours)	**36.82**	**44.49**	**68.28**	**71.87**	72.90	70.35	68.13	**48.32**	**46.21**	**41.22**
Joint Train	36.45	39.97	61.24	61.65	74.27	74.27	61.65	61.24	39.97	36.45

Table 1. Experimental results on five different attacks on CIFAR10.

Based on above analysis, a natural insight is that, in the popular 'pre-training & fine-tuning' training paradigm, downstream tasks may benefit from a strong robust pretraining. Conversely, a pre-trained model lacking robustness may face challenges in achieving robustness for downstream tasks. Furthermore, compared with standard continual learning, the sequence of attacks becomes a more important factor in continual defense. Models pre-trained with easy attacks may offer more generalizable representation for subsequent, more challenging attacks. Conversely, models adversarially trained with more challenging attacks may experience increased catastrophic forgetting but may obtain the shortcut to continuable robustness. Numerous intriguing properties in continuous defense still need to be further explored.

3. Expansion evaluation of AIR

Evaluation with more attacks. Our experiments followed the common evaluation setup in continual learning, typically involving sequences of 2-3 tasks. Additionally, our evaluation surpasses CAD[9] by considering sequence order and attack strengths, which CAD neglects. Still, we conduct additional experiments with a sequence length of 5

	FGSM & Patch				PGD_{L2} & $PGD_{L\infty}$				PGD & AA			
Tasks	FGSM to Patch		Patch to FGSM		L_2 to L_∞		L_∞ to L_2		PGD to AA		AA to PGD	
	Task1	Task2	Task1	Task2	Task1	Task2	Task1	Task2	Task1	Task2	Task1	Task2
Vanilla	21.31	59.55	17.19	**58.31**	64.41	43.09	21.88	77.65	16.84	74.51	64.10	41.31
AIR(ours)	**40.75**	**73.62**	**30.46**	54.48	**79.44**	**48.04**	**40.11**	**78.34**	**44.97**	**75.20**	**73.06**	**47.46**
Joint Train	43.83	57.76	57.76	43.83	78.39	44.84	44.84	78.39	40.89	74.50	74.50	40.89

Table 2. Experimental results of various attack strategies on CIFAR10.

Orders	PGD to FGSM			PGD to None		
Attacks	AA	Adv.Patch	SSAH	AA	Adv.Patch	SSAH
vanilla	0.17	11.48	13.58	0.08	23.58	10.30
AIR(ours)	**29.33**	**44.90**	**32.76**	**29.28**	**45.88**	**33.52**
AT_{PGD} (upper bound)	34.71	44.27	36.31	34.71	44.27	36.31

*

Table 3. Evaluation on additional attacks.

Tasks	FGSM to PGD		PGD to FGSM	
	Task1	Task2	Task1	Task2
vanilla	14.04	26.67	3.18	**30.04**
AIR(ours)	**22.62**	**28.01**	**21.07**	29.41
Joint train	30.76	28.34	28.34	30.76

Table 4. Exp. on TinyImageNet.

CIFAR10-C	None & Snow				GN & Contrast			
Tasks	None to Snow		Snow to None		GN to Contrast		Contrast to GN	
	Task1	Task2	Task1	Task2	Task1	Task2	Task1	Task2
vanilla	70.57	80.70	66.55	83.43	75.95	**87.32**	53.48	**87.96**
AIR(ours)	**78.41**	**84.21**	**77.48**	**84.25**	**79.64**	83.31	**76.58**	85.94
Joint train	84.34	82.10	82.10	84.34	86.36	84.39	84.39	86.36

Table 5. Experimental results on CIFAR-10-C dataset.

(with different attack strategies) on CIFAR10. The results in Table 1 show that our AIR also performs well in longer sequences with increased internal diversity.

Evaluation on AutoAttack (AA) Our initial focus was on establishing an explicit adaptation process for each attack, aligning with the common setup in continual learning, thereby excluding the evaluation of AA. Here, we conduct experiments to evaluate explicit adaptation (offline adversarial training) to attacks like AA and present the results in Table 2. We also provide the results in Table 3 as a demonstration of AIR's resistance to forgetting when facing attacks including AA, during adaptation to other attack sequence (e.g., PGD to None/FGSM). Our AIR exhibits stability under both settings.

Evaluation on ImageNet. We use CIFAR since it is the most common dataset for adversarial defense. we conduct an evaluation on Tiny-ImageNet as a lightweight proxy experiment and present the results in Table 4. Our AIR also demonstrates continual robustness on the larger datasets.

Evaluation on common corruptions. We appreciate your suggestion. Considering the diversity and combination of image corruptions, we conduct experiments to evaluate AIR's resistance to forgetting under three most challenging types of image corruptions, including *Snow*, *Gaussian_noise (GN)*, and *Contrast*, and present the results in Table 5. AIR also demonstrates stability against continual corruption tasks.

References

[1] 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. 1

[2] 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] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on computer vision and Pattern Recognition*, pages 770–778, 2016. 1

[4] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 2

[5] Hongshuo Liang, Erlu He, Yangyang Zhao, Zhe Jia, and Hao Li. Adversarial attack and defense: A survey. *Electronics*, 11(8):1283, 2022. 1

[6] 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

[7] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009. 2

[8] Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4322–4330, 2019. 1

[9] Qian Wang, Yaoyao Liu, Hefei Ling, Yingwei Li, Qihao Liu, Ping Li, Jiazhong Chen, Alan Yuille, and Ning Yu. Continual adversarial defense. *arXiv preprint arXiv:2312.09481*, 2023. 2

[10] 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. 1