Exploring Cross-Stage Adversarial Transferability in Class-Incremental Continual Learning

Jungwoo Kim, Jong-Seok Lee[†]

School of Integrated Technology / BK21 Graduate Program in Intelligent Semiconductor Technology
Yonsei University
Seoul, Republic of Korea
{kjungwoo, jong-seok.lee}@yonsei.ac.kr

Abstract—**Class-incremental continual learning addresses catastrophic forgetting by enabling classification models to preserve knowledge of previously learned classes while acquiring new ones. However, the vulnerability of the models against adversarial attacks during this process has not been investigated sufficiently. In this paper, we present the first exploration of vulnerability to stage-transferred attacks, i.e., an adversarial example generated using the model in an earlier stage is used to attack the model in a later stage. Our findings reveal that continual learning methods are highly susceptible to these attacks, raising a serious security issue. We explain this phenomenon through model similarity between stages and gradual robustness degradation. Additionally, we find that existing adversarial training-based defense methods are not sufficiently effective to stage-transferred attacks. Codes are available at https://github.com/mcml-official/CSAT.**

Index Terms—**continual learning, adversarial robustness, adversarial transferability.**

I. INTRODUCTION

Deep learning models trained on large-scale image datasets have achieved state-of-the-art performance in image classification. However, they face a critical challenge: adapting to new visual categories (i.e., classes) without losing previously learned knowledge. Unlike human cognitive flexibility, these models struggle with catastrophic forgetting, where learning new image classes rapidly erodes prior learning. Continual learning emerges as a promising approach to mimic human visual adaptability and overcome catastrophic forgetting. Specifically, class-incremental continual learning (Class-IL) [1] allows models to seamlessly integrate emerging visual categories while preserving existing knowledge. This capability is crucial for computer vision and multimedia systems operating in dynamic, ever-evolving environments, bridging the gap between artificial and human perception.

Meanwhile, ensuring model security and adversarial robustness has become a critical research topic nowadays. Traditional deep learning models have shown significant vulnerability to adversarial attacks, with existing studies [2]–[4] highlighting critical security challenges. The unique characteristics of Class-IL scenarios amplify these security risks. Models undergo iterative learning stages rather than being trained from scratch at once, creating potential vulnerabilities: even without direct access to the current stage model, adversaries can execute black-box attacks by leveraging models from earlier

training stages. This risk is significant due to the transferability of adversarial attacks [5]–[7] between different deep learning models. However, research investigating adversarial robustness within the Class-IL context remains relatively limited. Existing studies [8]–[13] are still in the early phases of understanding adversarial robustness degradation and to the best of our knowledge, no existing work has investigated *stage-transferred adversarial attacks* in Class-IL scenarios.

To address this research gap, we explore cross-stage adversarial robustness in Class-IL scenarios. To the best of our knowledge, our work is the first to address the effectiveness of perturbed images generated using the model weights at earlier stages in Class-IL. Our analysis shows that these stage-transferred attacks achieve comparable performance to direct attacks, revealing significant correlations between stage distances and transferability. We provide insights into the observed adversarial transferability patterns through comprehensive experiments of cross-stage model similarities and robustness degradation. Furthermore, we also evaluate the effectiveness of state-of-the-art adversarial-training-based defense strategies in mitigating these stage-transferred vulnerabilities.

II. RELATED WORKS

Continual learning is commonly organized along two axes: *(i)* the *scenario-based* and *(ii)* the *strategy-based*. Scenario-wise, three canonical paradigms are recognized [1]—*task-incremental*, *domain-incremental*, and the particularly demanding *class-incremental* (Class-IL) setting, in which the class label expands as training progresses. Strategy-wise, recent surveys divide the literature into replay-based and regularization-based methods [14]. The former approach such as incremental classifier and representation learning (iCaRL) [15] retains exemplars or synthetic memories to rehearse earlier knowledge, whereas the latter—e.g. elastic weight consolidation (EWC) [16] and learning without forgetting (LwF) [17]—penalizes parameter updates that would otherwise interfere with previous representations.

While standard continual learning addresses catastrophic forgetting, the security issue of continual learning has received attention only recently. After Khan *et al.* [8] first demonstrated the adversarial vulnerability of continual learning models, subsequent studies tried to improve adversarial robustness through adversarial training based on margin-aware loss [9],
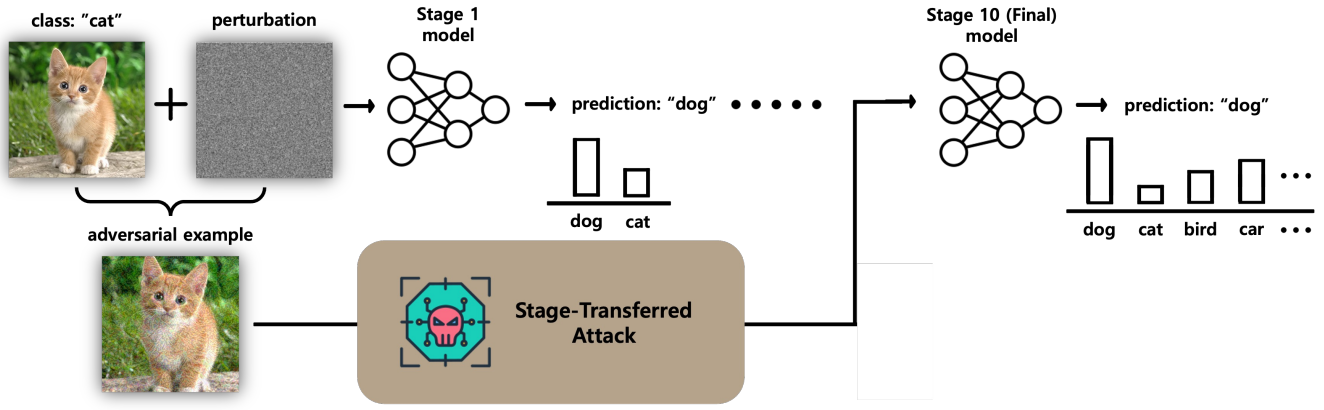
Fig. 1: **Framework of Stage-Transferred Attack.** Our framework for stage-transferred attack in class-incremental continual learning. The adversary crafts an adversarial example with perturbation generated in the earlier stage, and it is transferred to the final stage model.

boundary-preserving regularizers [10], or unified replay-attack optimization [13]. They implicitly assume that an attacker crafts perturbation with respect to the current model. On the other hand, the susceptibility of Class-IL models to *stage-transferred adversarial attacks* remains unexplored.

III. ADVERSARIAL TRANSFERABILITY

In Class-IL scenarios, a model is trained iteratively through multiple stages, where new classes to be identified are added in each stage. In this section, we explore stage-transferred attacks between these iterative stages. Specifically, we define *stage-transferred attacks* as adversarial examples crafted using model f_t at stage t, then applied to later stage model f_{t+k} where $k > 0$. Note that this differs from the general *adversarial transferability* between different deep learning models, as we focus on transfer across iteratively updated models within continual learning scenarios (Fig. 1).

A. Experimental Set-up

We design a Class-IL scenario using Avalanche [20], an end-to-end continual learning framework. We adopt several representative Class-IL approaches, including iCaRL [15], greedy sampler and dumb learner (GDumb) [18], experience replay with asymmetric cross-entropy (ER-ACE) [19], and experience replay with asymmetric metric learning (ER-AML) [19].

Our experiments evaluate these methods on two widely used benchmarks: Split-MNIST [21] and Split-CIFAR100 [22]. For Split-MNIST, we utilize multi-layer perceptron (MLP) models, while for Split-CIFAR100, we utilize a reconstructed ResNet-based architecture [23], reflecting the complexity of these datasets. The Split-MNIST benchmark is divided into five stages where two classes are added at each stage. The Split-CIFAR100 benchmark consists of ten stages with ten classes per stage. All hyper-parameter settings for the Class-IL methods, such as learning rate, memory buffer size, etc., are carefully configured to achieve performance comparable to the references.
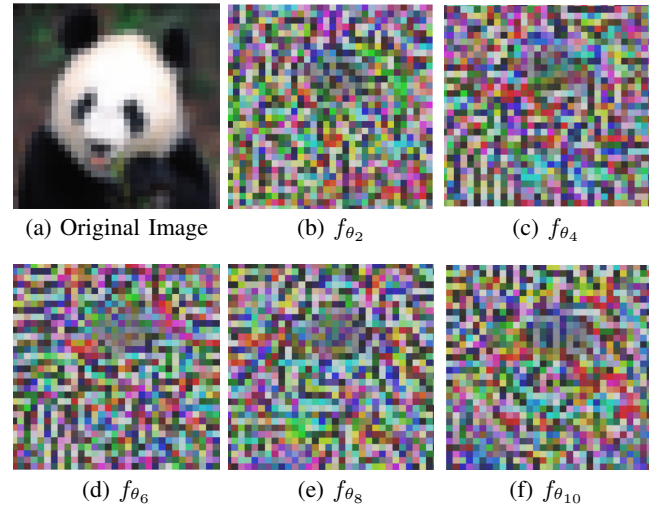

Fig. 2: **Adversarial Perturbation Visualization.** The leftmost panel shows the original image, while the rest shows the adversarial perturbations produced using each stage model (f_{θ_2} to $f_{\theta_{10}}$ for Split-CIFAR100). Pixel values have been rescaled to enhance visibility.

To evaluate stage-transferred attacks, we utilize three representative attack methods: fast gradient sign method (FGSM) [2], projected gradient descent (PGD) [3], and autoattack (AA) [4]. FGSM generates an adversarial example x' by applying a single-step perturbation in the direction of the sign of the gradient of the loss function $J(\theta, x, y)$ with respect to the input x, scaled by a perturbation magnitude ϵ:

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

Here, θ represents the model parameters, and y is the ground-truth label corresponding to x. $\nabla_x J(\theta, x, y)$ denotes the gradient of the loss function with respect to the input x. On the other hand, PGD iteratively updates the adversarial example. At each iteration τ, the perturbed input x^τ is updated in the direction of the sign of the gradient by a small step size α,

TABLE I: **Cross-stage Adversarial Transferability on the Split-MNIST Benchmark.** Each entry reports the ASR when adversarial examples—crafted by the attacker model of the indicated stage (column) using FGSM, PGD, or AA—are evaluated on the final-stage target model. Note that the last column corresponds to the case where the attacker and target models are the same, thus the results are from direct attacks to the target model.

Method	Attack	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5 (Final)
iCaRL [15]	FGSM	0.774	0.895	0.927	0.941	0.953
	PGD	0.774	0.895	0.927	0.941	0.953
	AA	0.774	0.841	0.878	0.751	0.953
GDumb [18]	FGSM	0.076	0.079	0.119	0.116	0.131
	PGD	0.072	0.078	0.114	0.110	0.127
	AA	0.005	0.019	0.027	0.048	0.150
ER-ACE [19]	FGSM	0.087	0.085	0.122	0.126	0.155
	PGD	0.078	0.080	0.112	0.119	0.149
	AA	0.004	0.012	0.054	0.065	0.188
ER-AML [19]	FGSM	0.020	0.038	0.093	0.129	0.141
	PGD	0.020	0.037	0.087	0.125	0.148
	AA	0.006	0.010	0.038	0.070	0.176

TABLE II: **Cross-stage Adversarial Transferability on the Split-CIFAR100 Benchmark.** Each entry reports the ASR when adversarial examples—crafted by the attacker model of the indicated stage (column) using FGSM, PGD, or AA—are evaluated on the final-stage target model. Note that the last column corresponds to the case where the attacker and target models are the same, thus the results are from direct attacks to the target model.

Method	Attack	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6	Stage 7	Stage 8	Stage 9	Stage 10 (Final)
iCaRL [15]	FGSM	0.697	0.704	0.743	0.747	0.773	0.787	0.802	0.802	0.822	0.836
	PGD	0.702	0.724	0.771	0.777	0.796	0.817	0.834	0.840	0.857	0.891
	AA	0.648	0.663	0.697	0.680	0.646	0.681	0.712	0.714	0.663	0.928
GDumb [18]	FGSM	0.701	0.728	0.677	0.696	0.708	0.733	0.717	0.713	0.972	0.802
	PGD	0.706	0.726	0.671	0.693	0.717	0.737	0.713	0.715	0.980	0.944
	AA	0.110	0.354	0.496	0.521	0.533	0.547	0.530	0.525	0.727	0.968
ER-ACE [19]	FGSM	0.836	0.853	0.746	0.774	0.806	0.824	0.769	0.778	0.793	0.795
	PGD	0.840	0.850	0.746	0.767	0.803	0.820	0.773	0.779	0.796	0.821
	AA	0.443	0.479	0.364	0.351	0.363	0.399	0.421	0.422	0.487	0.862
ER-AML [19]	FGSM	0.732	0.759	0.697	0.725	0.720	0.752	0.718	0.711	0.731	0.745
	PGD	0.736	0.775	0.702	0.720	0.719	0.748	0.721	0.712	0.737	0.783
	AA	0.273	0.353	0.314	0.279	0.327	0.328	0.398	0.381	0.375	0.835

followed by a projection $\Pi_{B_\epsilon(x)}$ to ensure that the perturbed sample remains within an ϵ-ball around the original input:

$$x^{\tau+1} = \Pi_{B_\epsilon(x)} \left(x^\tau + \alpha \cdot \text{sign}(\nabla_x J(\theta, x^\tau, y)) \right) \quad (2)$$

AA combines four strong and complementary attacks under the ℓ_∞ constraint: two adaptive projected gradient descents (APGD-CE and APGD-DLR), the fast adaptive boundary (FAB) attack [24], and the query-efficient square attack [25]. Given an input x, AA runs each sub-attack with fixed hyper-parameters and returns the first adversarial example that succeeds, thereby providing a deterministic and reproducible lower bound on model robustness without the need for manual tuning.

We set the perturbation magnitude as $\epsilon = 0.3$ for Split-MNIST and $\epsilon = 8/255$ for Split-CIFAR100. The number of iterations (K) is set 40 for Split-MNIST and 10 for Split-CIFAR100, and α in PGD is set to ϵ/K. Fig. 2 shows an example case of the generated perturbations using models of different stages.

B. Cross-Stage Adversarial Transferability

To evaluate the performance of stage-transferred attacks, we measure how effective adversarial examples crafted by the *attacker* model f_{θ_a} (*stage a*) are to the *target* model f_{θ_T}, where T indicates the last stage (5 for Split-MNIST and 10 for Split-CIFAR100). While all but the first stage can be the target of transferred attacks, we set the final stage as the target in order to highlight core results without nonessential complexity of analysis. We assume that the classes added in each stage are known to the adversary. Let $\mathcal{D}_{1:t}$ denote the cumulative test dataset from stage 1 to t.

Given a clean image sample x, the adversary uses θ_a to obtain the perturbed input x'_a. The attack success rate (ASR) of x'_a to f_{θ_T} is defined as follows:

$$\text{ASR}_{a \to T} = \frac{\sum_{(x,y) \in \mathcal{D}_{1:a}} \mathbb{1}\left[f_{\theta_T}(x) = y\right] \mathbb{1}\left[f_{\theta_T}(x'_a) \neq y\right]}{\sum_{(x,y) \in \mathcal{D}_{1:a}} \mathbb{1}\left[f_{\theta_T}(x) = y\right]} \quad (3)$$

Here, $f_\theta(\cdot)$ denotes the classification result of the model with parameters θ, and $\mathbb{1}[\cdot]$ is the indicator function. The numerator counts the number of samples that are *correctly*

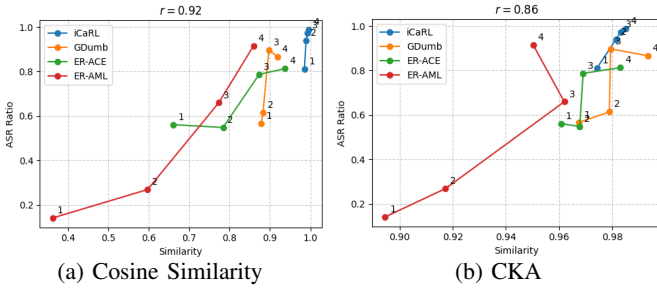(a) Cosine Similarity (b) CKA

Fig. 3: **Correlation between Model Similarity and Adversarial Transferability**. The number of each data point indicates the stage (a) whose model is compared to the final-stage model for Split-MNIST. Pearson correlation coefficients r between similarity and ASR ratio are also shown.

classified by the target model f_{θ_T} without perturbation (i.e., clean original images) but *misclassified* under stage-transferred attack. The denominator normalizes by the number of correctly classified clean images, yielding a stage-aware measure of adversarial vulnerability. Leveraging ASR, unlike simple classification accuracy, provides a more direct measure of adversarial transferability, as continual learning scenarios often involve performance degradation due to forgetting.

The results in Tables I and II show how adversarial examples generated using earlier stage models maintain their effectiveness on the final stage models in terms of ASR. It can be observed that in the Split-MNIST benchmark (Table I), the stage-transferred attacks are effective. For instance, in iCaRL, the transferred attack using FGSM from model f_{θ_1} achieves ASR of 0.774, which is only a slight drop by 0.179 from the ASR of direct attack on f_{θ_5} (0.953). This trend becomes more pronounced due to the larger number of classes and stages in Split-CIFAR100 (Table II). For instance, in GDumb, the transferred attack using FGSM from model f_{θ_1} remains nearly as effective as direct attack using $f_{\theta_{10}}$, showing ASR of 0.697 and 0.836, respectively.

Also, we can observe correlation between stage proximity, i.e. stage distance, and attack effectiveness: adversarial examples generated from later stages shows higher transferability compared to those from earlier stages. We further explore the underlying mechanisms behind this cross-stage adversarial transferability in the following sections.

C. Model Similarity

One factor causing cross-stage adversarial transferability shown in Tables I and II is the inter-stage model similarity. Several previous works [26]–[28] have shown the existence of significant correlation between adversarial transferability and model similarity. These findings have particularly concerning implications for Class-IL, as this scenario operates on a continuously evolving single model rather than separate, independent models. The inherent architectural similarity between stages likely amplifies the model's vulnerability to transferred adversarial attacks, potentially making these models more

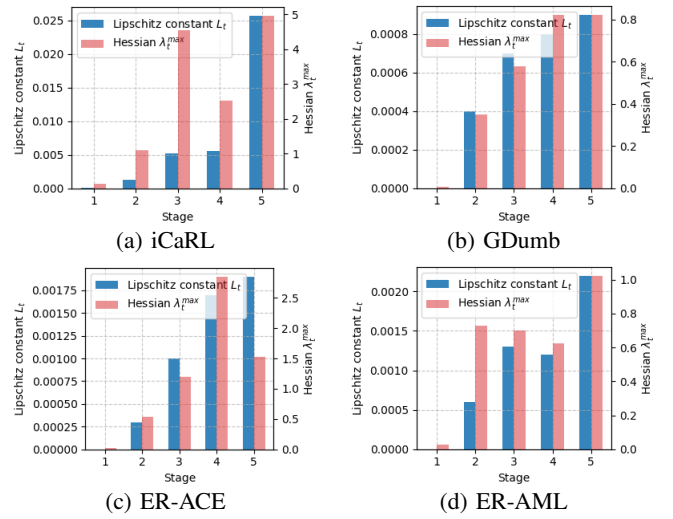(a) iCaRL (b) GDumb

(c) ER-ACE (d) ER-AML

Fig. 4: **Stage-wise Model Complexity.** Stage-wise growth of the Lipschitz constant and the spectral norm of the Hessian on the Split-MNIST benchmark. Higher values indicate higher complexity.

TABLE III: **Asymmetric Transferability.** ASR of PGD for perturbations crafted on the earlier model and tested on the final model ($4 \rightarrow 5$) or the reverse ($5 \rightarrow 4$) on the Split-MNIST benchmark.

Attacker $\rightarrow$ Target	iCaRL	GDumb	ER-ACE	ER-AML
$4 \rightarrow 5$	0.941	0.109	0.118	0.127
$5 \rightarrow 4$	0.843	0.047	0.108	0.094

susceptible to attacks than traditional, independently trained models.

To quantify the impact of cross-stage model similarity on adversarial transferability, we quantitatively measure the similarity between an early-stage model and the final-stage model in terms of cosine similarity between the parameters of the two models [29] and centered kernel alignment (CKA) [30]. Here, CKA assesses representation-level similarity by correlating the centered Gram matrices of activations from the two models, yielding a scale- and rotation-invariant measure of how closely their feature spaces align. Both metrics are computed using the data for the task of the first stage ($\mathcal{D}_1$), which is commonly learned across all stage models.

Fig. 3 plots the similarity vs. adversarial transferability. The relative adversarial transferability is measured as the ASR ratio, which is defined as $\text{ASR}_{a \rightarrow T}$ normalized by ASR of the direct attack, i.e., $\text{ASR}_{a \rightarrow T} / \text{ASR}_{T \rightarrow T}$. As a increases, i.e., the stage of the attacker model is closer to the final stage, their similarity increases and consequently the adversarial transferability also increases. This trend is consistent for both similarity metrics. This result explains the correlation between stage proximity and adversarial transferability observed in Section III-B.

D. Model Complexity and Robustness Degradation

Model similarity is not the only source of cross-stage adversarial transferability. If similarity is the sole driver, adversarial perturbations crafted in the later model (e.g., f_{θ_5}) should succeed just as well on the earlier model (e.g., f_{θ_4}), because the pairwise similarities $\texttt{sim}(\theta_4, \theta_5)$ and $\texttt{sim}(\theta_5, \theta_4)$ are identical by definition. Empirically, however, the transferability is asymmetric: Table III shows that the backward transferability $(5 \rightarrow 4)$ is lower than the forward transferability $(4 \rightarrow 5)$. This implies that there are other factors contributing to cross-stage adversarial transferability observed in Section III-B than model similarity. In particular, we identify gradual robustness degradation due to increasing model complexity during the Class-IL process.

Previous studies [8], [12] have shown that, in continual learning, model robustness tends to decrease as more stages are introduced and the number of classes increases. To isolate this effect, we quantify how each stage's decision boundary becomes complex and thus vulnerable to attack. Concretely, for each stage model f_{θ_t}, we compute two complexity metrics. One is the average local Lipschitz constant [31] defined as

$$L_t = \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla_x J(\theta_t, x_i, y_i) \right\|_2 \qquad (4)$$

where N is the number of data. The other is the spectral norm of the Hessian matrix of the loss with respect to the weights W of the last layer [32], i.e.,

$$\lambda_t^{\max} = \lambda_{\max}\left(\nabla_W^2 J(\theta_t, x_i, y_i) \right) \qquad (5)$$

where $\lambda_{\max}$ is the maximum singular value. The former captures how rapidly the loss changes in the input space, while the latter reflects the curvature of the decision boundary in the parameter space. Fig. 4 shows that L_t and $\lambda_t^{\max}$ grow as the stage progresses. Intuitively, as new classes are added, the model's feature space changes to fit more finer-grained boundaries, making them steeper and more complex.

Taken together, these observations indicate a two-step mechanism behind cross-stage adversarial transferability in Class-IL. First, due to model similarity between stages, models from different stages share the direction along which adversarial perturbation can cross the decision boundary and induce misclassification. Second, as each new stage carves finer decision boundaries to accommodate an increased number of classes, the loss surface becomes progressively steeper, making the target model in a later stage vulnerable to the shared perturbation direction.

IV. Defense against Stage-Transferred Attacks

Building on our analysis of cross-stage adversarial transferability in Section III, we now explore the possibility of defending against such stage-transferred attacks.

A. Defense Strategies

Adversarial training is a common approach to improve model robustness [2]. However, in Class-IL scenarios, the robustness acquired in an earlier stage may be forgotten in later stages. Therefore, adversarial training adapted to Class-IL has been proposed. We test three representative methodologies: Task aware boundary augmentation (TABA) [9], flatness-preserving adversarial incremental learning for robustness (FLAIR) [13], and FLAIR+ [13]. TABA extracts boundary buffers with PGD-misclassified samples and leverages them with task-aware mixup. FLAIR integrates separated logit-based knowledge distillation and flatness-preserving distillation into the PGD process, and FLAIR+ trains FLAIR with RandAugment [33].

B. Defense with Stage-Transferred Attack

Fig. 5 shows the results of applying the three defense strategies to the stage-transferred attacks on Split-CIFAR100. While ASR is lowered to some extent by the three defense methods in comparison to the results in Table II, they still remain vulnerable to the stage-trasferred attacks. For instance, FGSM shows nearly 50% of ASR when f_{θ_9} is used as the attacker model. Combined with the high stage-transferred ASR observed in Section III, these results underscore an urgent need for defense strategies specifically tailored to the characteristics and vulnerabilities of Class-IL scenarios.

V. Conclusion

In this work, we have presented our investigation of adversarial vulnerability to stage-transferred attacks in Class-IL. We showed that stage-transferred attacks are highly transferable across stages, posing critical security issues in Class-IL scenarios. Model similarity and robustness degradation were shown to be main causes of the cross-stage adversarial transferability. Additionally, we found that existing adversarial training-based defense strategies are not sufficient to mitigate these vulnerabilities. Our findings highlight the need for future research to develop effective defense mechanisms against stage-transferred attacks in continual learning.

We focused on stage-transferred adversarial attacks in Class-IL. Yet, real-world deployments may encounter multiple threat vectors; data poisoning, exploratory black-box attacks, and back-door insertions may co-occur and further erode the security posture of continual learners. Future works should integrate these various attack scenarios into a unified evaluation framework and quantify how their interactions amplify cumulative vulnerabilities.

(a) TABA [9]   (b) FLAIR [13]   (c) FLAIR+ [13]
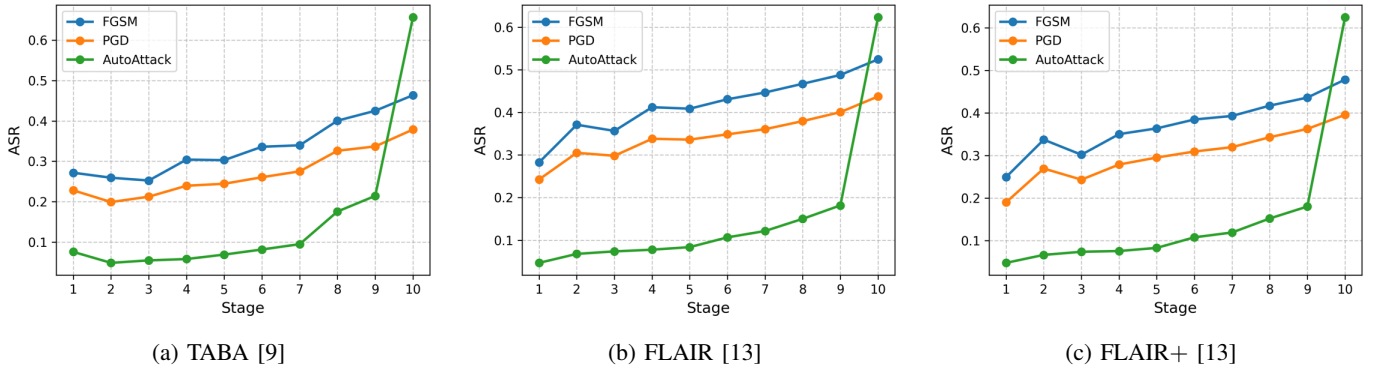
Fig. 5: **Defense Methods against Cross-Stage Adversarial Transferability.** Each panel reports the ASR when adversarial examples are crafted at the indicated stage with FGSM, PGD, or AutoAttack and evaluated on the final-stage target model $(f_{\theta_{10}})$ on the Split-CIFAR100 benchmark.

REFERENCES

[1] G. M. van de Ven, T. Tuytelaars, and A. S. Tolias, "Three types of incremental learning," *Nature Machine Intelligence*, vol. 4, pp. 1185–1197, 2022.

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[4] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2020.

[5] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.

[6] W. Wu, Y. Su, M. R. Lyu, and I. King, "Improving the transferability of adversarial samples with adversarial transformations," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] J. Gu, X. Jia, P. de Jorge, W. Yu, X. Liu, A. Ma, Y. Xun, A. Hu, A. Khakzar, Z. Li, X. Cao, and P. Torr, "A survey on transferability of adversarial examples across deep neural networks," *Trans. Mach. Learn. Res. (TMLR)*, 2024.

[8] H. Khan, P. M. Shah, S. F. A. Zaidi, Q. Zia, et al., "Susceptibility of continual learning against adversarial attacks," *arXiv preprint arXiv:2207.05225*, 2023.

[9] T. Bai, C. Chen, L. Lyu, J. Zhao, and B. Wen, "Towards adversarially robust continual learning," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[10] K. Mukai, S. Kumano, N. Michel, L. Xiao, and T. Yamasaki, "Adversarially robust continual learning with anti-forgetting loss," in *Proc. Int. Conf. Image Processing (ICIP)*, 2024.

[11] M. Umer and R. Polikar, "Adversary aware continual learning," *IEEE Access*, vol. 12, pp. 126108–126121, 2024.

[12] X. Ru, X. Cao, Z. Liu, J. M. Moore, X.-Y. Zhang, X. Zhu, W. Wei, and G. Yan, "Maintaining adversarial robustness in continuous learning," *arXiv preprint arXiv:2402.11196*, 2024.

[13] S. Cho, H. Lee, and C. Kim, "Enhancing robustness in incremental learning with adversarial training," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, 2025.

[14] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5362–5383, 2024.

[15] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Natl. Acad. Sci. (PNAS)*, 2017.

[17] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, 2017.

[18] A. Prabhu, P. H. S. Torr, and P. K. Dokania, "Gdumb: A simple approach that questions our progress in continual learning," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2020.

[19] L. Caccia, R. Aljundi, N. Asadi, T. Tuytelaars, J. Pineau, and E. Belilovsky, "New insights on reducing abrupt representation change in online continual learning," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022.

[20] A. Carta, L. Pellegrini, A. Cossu, H. Hemati, and V. Lomonaco, "Avalanche: A pytorch library for deep continual learning," *J. Mach. Learn. Res. (JMLR)*, vol. 24, no. 363, pp. 1–6, 2023.

[21] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, 2012.

[22] Alex Krizhevsky, "Learning multiple layers of features from tiny images," Master's thesis, Department of Computer Science, University of Toronto, 2009.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," in *Proc. Int. Conf. Machine Learning (ICML)*, 2020.

[25] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2020.

[26] J. Hwang, D. Han, B. Heo, S. Park, S. Chun, and J.-S. Lee, "Similarity of neural architectures based on input gradient transferability," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2024.

[27] F. Waseda, S. Nishikawa, T.-N. Le, H. H. Nguyen, and I. Echizen, "Closer look at the transferability of adversarial examples: How they fool different models differently," in *Proc. Int. Conf. Applications of Computer Vision (WACV)*, 2023.

[28] G. Klause and N. Bunzel, "The relationship between network similarity and transferability of adversarial attacks," *arXiv preprint arXiv:2501.18629*, 2025.

[29] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[30] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019.

[31] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel, "Evaluating the robustness of neural networks: An extreme value theory approach," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[32] M. Sabanayagam, F. Behrens, U. Adomaityte, and A. Dawid, "Unveiling the hessian's connection to the decision boundary," *arXiv preprint arXiv:2306.07104*, 2023.

[33] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. Adv. Neural Information Processing Systems (NeurIPS)*, 2020.