Robustness of Fuzzy ARTMAP to Adversarial Attacks and Progressive Adversarial Training for Streaming Learning

Shane Cairns[†], Leonardo Enzo Brito da Silva[‡], Sasha Petrenko[†], Donald C. Wunsch II[*†], and Jian Liu[*]
[*]Kummer Institute Center for Artificial Intelligence and Autonomous Systems (KICAIAS),
Missouri University of Science and Technology, Rolla, MO, USA
[†]Applied Computational Intelligence Lab (ACIL), Department of Electrical and Computer Engineering,
Missouri University of Science and Technology, Rolla, MO, USA
[‡]Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte,
Natal, RN 59078-900, Brazil

Email: sacg3q@mst.edu, leonardo.brito@imd.ufrn.br, sasha.petrenko@mst.edu, wunsch@mst.edu, jliu@mst.edu

***Abstract*—Incremental learners deployed on streaming data must remain robust to evolving adversarial perturbations, yet most adversarial-robustness studies assume offline multi-epoch training with repeated access to historical data. We investigate adversarial robustness in Fuzzy ARTMAP, a prototype-based Adaptive Resonance Theory (ART) model that supports single-pass learning without replay. To enable strong adaptive evaluation despite ARTMAP's non-differentiable winner-take-all dynamics, we propose WB-Softmax, a differentiable softmax relaxation that aggregates category-level activations into class-level scores for gradient-based attacks. Across USPS, MNIST, and Fashion-MNIST, WB-Softmax Projected Gradient Descent (PGD) achieves 89–100% attack success on vanilla models, exceeding transfer and query-based baselines at matched budgets and satisfying anti-gradient-masking sanity checks. We then study adversarial training under true streaming constraints by comparing offline versus online adversarial example generation and standard versus selective updates. Offline adversarial training consistently collapses robustness despite substantial category growth, whereas online training is dataset-dependent. We introduce progressive two-stage selective training that combines selective filtering with ϵ scheduling, achieving the best overall robustness (AURAC) on all three datasets (28.2%, 64.5%, and 41.3%). Finally, leveraging ART's explicit category geometry, we monitor incremental cluster validity indices to diagnose separation collapse and design a separation-aware absorption rule that improves high-ϵ robustness on USPS (11.0$\pm$1.4% at ϵ=0.30).**

***Index Terms*—Adaptive Resonance Theory, Fuzzy ARTMAP, adversarial robustness, adversarial training, incremental learning**

I. Introduction

In deployed settings, models must learn incrementally while remaining robust to evolving adversarial inputs (e.g., autonomous systems, fraud detection, and streaming medical diagnosis). However, most adversarial robustness research targets batch-trained deep neural networks using attacks such as the Fast Gradient Sign Method (FGSM) [1] and Projected Gradient Descent (PGD) [2], and does not address truly incremental learning systems where data arrives sequentially and retraining from scratch is infeasible [3], creating a critical gap because many defenses assume storing/replaying historical data that may be unavailable due to privacy or poisoning risks.

Since adversarial examples were first documented [4], a large body of work has developed attack methods and robustness evaluation protocols [1]–[3], and pitfalls such as gradient masking motivate the sanity checks we adopt [5]. RobustBench standardizes adversarial evaluation practices [6], and recent surveys summarize adversarial training and adversarial ML more broadly [7]. Nevertheless, this literature focuses almost exclusively on deep networks and typically assumes offline, multi-epoch training with repeated access to historical data, which is often incompatible with true single-pass streaming.

Recent work has begun examining the intersection of continual learning and adversarial robustness [8], [9]. For example, Mi *et al.* [10] study robustness in memory-based continual learning and show that adversarial samples can accelerate catastrophic forgetting, while related approaches address robustness under non-stationary/online settings [11]. However, many such methods rely on replay buffers or regularization mechanisms that remain difficult to reconcile with strict single-pass streaming, where each example is processed once and then discarded. This mismatch motivates studying robustness in architectures that natively support incremental learning without replay.

Adaptive Resonance Theory (ART) networks [12], [13] provide a principled solution to this challenge. Unlike standard neural networks that suffer from catastrophic forgetting when trained on new data [8], ART supports single-pass incremental learning without requiring replay of previous samples. ART's decision boundaries correspond to explicit category regions defined by weight vectors, yielding a prototype-based representation with interpretability advantages over traditional deep networks. The vigilance mechanism provides built-in control over category granularity and can serve as a novelty detection signal for out-of-distribution or adversarial inputs. These

properties make ART attractive for online deployment, and motivate understanding and strengthening ART's adversarial robustness before using it in sensitive environments.

Fuzzy ARTMAP [14], the supervised learning variant of ART, enables incremental learning with explicit prototype-based categories. Recent work has developed modular ART implementations [15], deep hierarchical extensions [16], and gradient-free deep learning approaches using ART dynamics [17]; analysis of match-tracking mechanisms further reveals computational trade-offs across ARTMAP variants [18]. To our knowledge, no prior work addresses ARTMAP's adversarial robustness. This gap is not merely empirical: core ARTMAP operations include winner-take-all category selection and piecewise computations, which complicate standard gradient-based attack design and can lead to misleading conclusions without adaptive, mechanism-aligned evaluation.

This paper investigates four questions: (i) how fragile vanilla Fuzzy ARTMAP is under strong adaptive white-box attacks aligned with winner-take-all competition and complement coding; (ii) how adversarial training outcomes depend on training protocol (offline vs. online generation and selective vs. non-selective use of adversarial samples); (iii) whether progressive training via perturbation scheduling and selective filtering yields consistent robustness gains across datasets; and (iv) whether ARTMAP's explicit category geometry enables structural monitoring to identify failure modes and motivate targeted mitigations at high perturbation strengths.

We make four main contributions[1]. First, we develop WB-Softmax, an ART-aligned white-box attack using a differentiable relaxation of winner-take-all competition, achieving 89–100% attack success. Second, we show that adversarial-training outcomes depend critically on protocol: offline training collapses robustness despite category growth, whereas online training is dataset-dependent. Third, we introduce progressive two-stage selective training, combining selective filtering with epsilon scheduling, which achieves the highest AURAC across USPS, MNIST, and Fashion-MNIST (28.2%, 64.5%, and 41.3%). Fourth, across three datasets and three seeds, we use explicit category geometry and iCVIs to identify separation collapse and develop separation-aware training, improving high-ϵ USPS robustness from 5.4±0.6% to 11.0±1.4% at ϵ=0.30.

The remainder of the paper reviews preliminaries and the threat model, then presents WB-Softmax, interpretable training diagnostics, the experimental setup, results, and conclusions.

II. Preliminaries

A. Background

Fuzzy ART [19] extends Adaptive Resonance Theory to continuous inputs using fuzzy set operations. Given $\boldsymbol{x} \in [0,1]^d$, *complement coding* forms $\boldsymbol{I}(\boldsymbol{x}) = [\boldsymbol{x}; \mathbf{1}-\boldsymbol{x}] \in [0,1]^{2d}$ so that $|\boldsymbol{I}| = d$ is constant, which helps mitigate category proliferation due to input-norm variation. Here $|\boldsymbol{v}| = \|\boldsymbol{v}\|_1 = \sum_i v_i$ for $\boldsymbol{v} \in [0,1]^m$, and thus $|\boldsymbol{I}(\boldsymbol{x})| = \sum_{i=1}^d x_i + \sum_{i=1}^d (1-x_i) = d$.

Each category j has a weight vector $\boldsymbol{w}_j \in [0,1]^{2d}$ that defines a hyperbox in the coded space. The *match function* is

$$M_j(\boldsymbol{I}) = \frac{|\boldsymbol{I} \wedge \boldsymbol{w}_j|}{|\boldsymbol{I}|}, \tag{1}$$

and the *choice function* governing competition is

$$T_j(\boldsymbol{I}) = \frac{|\boldsymbol{I} \wedge \boldsymbol{w}_j|}{\alpha + |\boldsymbol{w}_j|}, \tag{2}$$

where $\wedge$ is element-wise minimum and $|\cdot|$ denotes the L^1 norm. The winner $J = \arg\max_j T_j(\boldsymbol{I})$ is accepted if $M_J(\boldsymbol{I}) \geq \rho$, where vigilance $\rho \in [0,1]$ controls category granularity (higher ρ yields finer categories). If vigilance fails, mismatch reset inhibits J and the search continues; when a category is accepted, fast learning updates $\boldsymbol{w}_J^{\text{new}} = \boldsymbol{I} \wedge \boldsymbol{w}_J^{\text{old}}$.

Fuzzy ARTMAP [14] couples an input module ART_a with a label module ART_b through a map field F^{ab} that associates input categories with class labels. With map-field vigilance $\rho_{ab} = 1.0$, each category maps to exactly one label. When a prediction error occurs, *match tracking* raises ρ_a minimally to force selection (or creation) of a more specific category that predicts the correct label.

ARTMAP's winner-take-all selection and piecewise operations complicate direct application of standard gradient-based white-box attacks; therefore, we use an ART-aligned differentiable objective (WB-Softmax) for adaptive evaluation (Section III).

B. Threat Model and Evaluation Protocol

We follow established best practices [3], [6] to formalize threat models and to distinguish genuine robustness from gradient masking artifacts [5]. In the *white-box* setting, the attacker has full access to model internals (weights $\{\boldsymbol{w}_j\}$, map field F^{ab}, and hyperparameters). In the *black-box* setting, the attacker has no internal access and must rely on transfer from a surrogate or gradient-free query methods. We assume the attacker has access to the training distribution and ground-truth labels, but not ARTMAP internals, and thus trains labeled surrogates for transfer attacks.

We consider untargeted ℓ_∞ attacks in normalized pixel space $[0,1]$ with budgets $\epsilon \in \{0.05, 0.10, \ldots, 0.35\}$, enforcing $\boldsymbol{x}_{\text{adv}} \in [0,1]^d$. For gradient-based attacks, FGSM [1] uses a single step, $\boldsymbol{x}_{\text{adv}} = \text{clip}_{[0,1]}\big(\boldsymbol{x} + \epsilon\,\text{sign}(\nabla_{\boldsymbol{x}}\mathcal{L})\big)$, and PGD [2] iterates this update with step size $\eta = \epsilon/4$ and projects back onto the ℓ_∞ ϵ-ball after each step, where $\mathcal{L}$ denotes the attack loss defined in Section III.

Because ARTMAP includes non-differentiable winner selection and piecewise operations, our primary white-box evaluation uses the ART-aligned WB-Softmax PGD objective (defined in Section III); complementary black-box baselines include transfer PGD from CNN surrogates (including LRS-enhanced surrogates [20]); square attack is used as an additional gradient-free baseline on vanilla models (Table V).

[1]Code available at: `https://github.com/syre-ai/adv-art`

We report clean accuracy, robust accuracy as a function of ϵ, and the Area Under the Robust Accuracy Curve (AURAC),

$$\text{AURAC} = \frac{1}{\epsilon_{\max} - \epsilon_{\min}} \int_{\epsilon_{\min}}^{\epsilon_{\max}} \text{Acc}(\epsilon)\, d\epsilon, \tag{3}$$

computed by trapezoidal integration over $\epsilon \in [0.05, 0.35]$, where $\text{Acc}(\epsilon)$ denotes robust accuracy on the clean-correct subset unless stated otherwise.

Following Athalye *et al.* [5], we apply four sanity checks adapted for ART: (i) FGSM achieves non-trivial attack success, (ii) iterative PGD is at least as strong as FGSM, (iii) white-box attacks match or exceed black-box transfer, and (iv) accuracy degrades smoothly with increasing ϵ. These checks help rule out gradient masking and suboptimal attack configurations.

III. White-Box Attacks on Fuzzy ARTMAP

We introduce WB-Softmax, an ART-aligned differentiable objective that enables adaptive white-box evaluation despite ARTMAP's non-differentiable dynamics.

WB-Softmax via softmax-relaxed prediction loss. ART's winner-take-all competition selects the category with highest choice function $J = \arg\max_j T_j(\boldsymbol{I})$, then predicts through the map field. The argmax has zero gradient almost everywhere, preventing direct optimization. Our key insight is to replace the hard winner selection with a differentiable softmax relaxation that provides a strong surrogate loss for optimization. Algorithm 1 summarizes the complete WB-Softmax PGD procedure used throughout this paper. We use $\mathcal{U}(-\epsilon, \epsilon)$ to denote element-wise uniform noise in $[-\epsilon, \epsilon]$. $\Pi_{[\boldsymbol{x}-\epsilon, \boldsymbol{x}+\epsilon]}(\cdot)$ denotes projection onto the ℓ_∞ ball around $\boldsymbol{x}$, implemented by coordinate-wise clipping to $[\boldsymbol{x} - \epsilon, \boldsymbol{x} + \epsilon]$.

The WB-Softmax loss computes class *probabilities* via softmax attention over category choice values. Let $\mathcal{C}_c$ denote the set of categories mapping to class c. We apply softmax over all category activations and aggregate within each class:

$$p_c = \sum_{j \in \mathcal{C}_c} \frac{\exp(T_j/\tau)}{\sum_k \exp(T_k/\tau)}, \tag{4}$$

where τ controls the softmax temperature. The attack loss is the negative log-likelihood of the true class:

$$\mathcal{L}_{\text{WB-Softmax}} = -\log(p_y). \tag{5}$$

We set $\tau = 0.01$; lower temperatures concentrate probability mass on winning categories while retaining gradient signal. Maximizing this loss pushes probability mass away from the true class, promoting misclassification. Eq. (4) aligns the attack objective with ARTMAP's competition-and-mapping behavior by converting category-level choice values into class-level scores consistent with the map field, rather than attacking an unrelated proxy. This confirms that effective white-box attacks exist when the objective captures ART's competition-and-mapping behavior.

Smooth surrogates and complement coding. WB-Softmax requires smooth approximations of ART's non-differentiable operations. We introduce differentiability only in the aggregation step (softmax over categories), keeping the forward ARTMAP computations intact (unlike BPDA-style gradient substitutions [5]). This design keeps the forward computation faithful to ARTMAP while introducing differentiability only at the aggregation stage used to construct the white-box objective.

Algorithm 1 WB-Softmax PGD Attack

Require: $\boldsymbol{x} \in [0,1]^d$, label y, budget ϵ, steps K, step size $\eta{=}\epsilon/4$, temperature τ

$\boldsymbol{x}' \leftarrow \text{clip}_{[0,1]}\big(\boldsymbol{x} + \mathcal{U}(-\epsilon, \epsilon)\big)$
for $k = 1$ **to** K **do**
  $\boldsymbol{I} \leftarrow [\boldsymbol{x}'; \mathbf{1} - \boldsymbol{x}']$
  $T_j \leftarrow \frac{|\boldsymbol{I} \wedge \boldsymbol{w}_j|}{\alpha_{\text{ART}} + |\boldsymbol{w}_j|} \;\; \forall j$
  $p_c \leftarrow \sum_{j \in \mathcal{C}_c} \text{softmax}(\boldsymbol{T}/\tau)_j \;\; \forall c$ ▷ class probabilities
  $\mathcal{L} \leftarrow -\log(p_y)$
  $\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{x}'} \mathcal{L}$ ▷ autograd (equiv. to Eq. (6))
  $\boldsymbol{x}' \leftarrow \Pi_{[\boldsymbol{x}-\epsilon, \boldsymbol{x}+\epsilon]}\big(\boldsymbol{x}' + \eta\, \text{sign}(\boldsymbol{g})\big)$; $\boldsymbol{x}' \leftarrow \text{clip}_{[0,1]}(\boldsymbol{x}')$
end for
return $\boldsymbol{x}'$

Given complement coding $\boldsymbol{I}(\boldsymbol{x}) = [\boldsymbol{x}; \mathbf{1} - \boldsymbol{x}]$, the chain rule yields:

$$\nabla_{\boldsymbol{x}} \mathcal{L} = \nabla_{\boldsymbol{I}_{1:d}} \mathcal{L} - \nabla_{\boldsymbol{I}_{d+1:2d}} \mathcal{L} \tag{6}$$

Eq. (6) ensures gradients are computed correctly under complement coding, since perturbations in $\boldsymbol{x}$ induce opposite-signed changes in the second half of $\boldsymbol{I}(\boldsymbol{x})$.

Black-box transfer attacks. For black-box evaluation, we train a SimpleCNN surrogate on the same data and generate adversarial examples using PGD (20 steps, step size $\epsilon/4$). These transfer to Fuzzy ARTMAP despite architectural differences, confirming cross-architecture vulnerability [21]. For stronger transfer attacks, we impose Lipschitz regularization on the loss landscape of our SimpleCNN for better transferability [20]. The LRS surrogate is trained for 10 epochs (sufficient for strong surrogate accuracy on these datasets) using Adam (lr $= 10^{-3}$) with gradient norm regularization $\lambda = 2500$. We use a single value $\lambda = 2500$ across all datasets, selected empirically from a brief sweep ($\lambda \in \{500, 1000, 2000, 2500\}$), and observed that black-box attack strength (AURAC) is relatively insensitive to λ in this range. These transfer attacks are strongest on USPS and remain weaker than WB-Softmax overall, providing complementary black-box baselines that help rule out gradient masking.

IV. Interpretable Diagnostics and Training Rules

ART exposes its internal state as explicit category geometry (hyperboxes), which enables a diagnosis-to-rule workflow that is difficult to realize in black-box models: we (i) monitor structural pathologies that emerge during adversarial learning, and (ii) translate them into targeted training rules that intervene directly on category formation and expansion. We demonstrate this workflow on USPS, where geometry-based evidence of *separation collapse* motivates a separation-aware absorption rule that improves high-ϵ robustness. Importantly, both the diagnostics and the training rules operate on the current category set and do not require replay buffers, preserving the single-pass streaming setting.

Algorithm 2 Adversarial Training Protocols (Offline vs. Online; Standard vs. Selective)

Offline (standard or selective):
  Train on all clean samples $\{(\boldsymbol{x}_i, y_i)\}$
  for each $(\boldsymbol{x}_i, y_i)$ **do**
    Generate $\boldsymbol{x}_i'$ against fixed model
    if standard **or** $f(\boldsymbol{x}_i') \neq y_i$: train on $(\boldsymbol{x}_i', y_i)$

Online (standard or selective):
  for each $(\boldsymbol{x}_i, y_i)$ **do**
    Train on $(\boldsymbol{x}_i, y_i)$
    Generate $\boldsymbol{x}_i'$ against *current* model
    if standard **or** $f(\boldsymbol{x}_i') \neq y_i$: train on $(\boldsymbol{x}_i', y_i)$

A. Online Structural Monitoring via iCVIs and Geometry Indicators

Incremental cluster validity indices (iCVIs) are online counterparts of classic cluster-quality measures that can be updated recursively as samples arrive [22], [23]. Unlike deep neural networks—where internal representations often require post-hoc embedding analysis—ART categories are explicit geometric objects, enabling direct computation of class-conditional separation/overlap/compactness metrics during training.

In this work, we monitor two geometry indicators that are especially informative during adversarial training: (i) *minimum separation*, defined as the smallest non-overlap margin between any pair of different-class categories; and (ii) *overlap risk*, defined as the maximum normalized intersection between wrong-class category pairs. Because these indicators depend only on the current category weights $\{\boldsymbol{w}_j\}$, they can be computed online without storing past samples. As shown in our USPS case study, these indicators reveal failure modes that may not be visible from accuracy curves or category counts alone, and they provide actionable signals for designing targeted training rules.

B. Streaming Adversarial Training Protocols, Failure Modes, and Targeted Rules

We compare adversarial training protocols along two dimensions: *when* adversarial examples are generated (offline vs. online) and *which* adversarial examples are used for updating (standard vs. selective). In offline training, adversarial examples are generated against a fixed clean-trained model and then used in a second pass. In online training, each clean sample is immediately followed by an adversarial example generated against the *current* model state. Standard variants train on all generated adversarial examples, whereas selective variants update only on adversarial examples that induce misclassification ($f(\boldsymbol{x}') \neq y$). Algorithm 2 summarizes the offline/online and standard/selective adversarial training protocols. Here, offline and online variants differ only in whether $\boldsymbol{x}'$ is generated against a fixed clean-trained model or the current model state.

Adversarial training can cause category proliferation, creating $\sim 2\times$ more categories, which increases memory footprint and evaluation cost. Intuitively, when an adversarial example $\boldsymbol{x}'$ is presented, perturbations may push it outside existing hyperboxes, triggering mismatch reset and category creation rather than absorption. Selective updating mitigates unnecessary growth by focusing learning on genuinely failing adversarial examples.

TABLE I: Separation Collapse on USPS ($\rho_a = 0.9$). Cats = number of categories; Min Sep = minimum separation; Compact = compactness index (higher is tighter)

Stage	Cats	Min Sep	Compact
Before (clean only)	6,058	0.036	0.994
After (selective adv.)	11,434	0.008	0.998
Change	$+1.9\times$	-78%	$+0.4\%$

To stabilize learning across attack strengths, we employ two-stage selective training with progressive ϵ scheduling: Stage 1 samples $\epsilon \in [0.05, 0.15]$ and Stage 2 extends to $\epsilon \in [0.15, 0.35]$, using online selective filtering in both stages. This schedule separates "learning to resist moderate attacks" from "adapting to stronger attacks," reducing instability that can occur when training is exposed immediately to large ϵ.

USPS case study: separation collapse. Geometry monitoring reveals a distinct failure mode under selective adversarial training: *separation collapse*. As categories expand to absorb adversarial examples, they may encroach upon regions associated with other classes, reducing inter-class margins even when categories remain locally compact. We quantify this effect via minimum separation. Table I shows that, on USPS, minimum separation drops by 78% during selective training, while compactness remains nearly unchanged. This indicates that categories remain tight but expand toward wrong-class regions, and that category count alone is insufficient to characterize robustness.

Separation-aware training rule. Motivated by separation collapse, we introduce separation-aware training, which modifies the absorption decision for adversarial examples. For each adversarial example $\boldsymbol{x}'$, we (i) identify the best-matching correct-class category j, (ii) simulate the post-update hyperbox $\boldsymbol{w}_j^{\text{new}} = \boldsymbol{w}_j \wedge \boldsymbol{I}(\boldsymbol{x}')$, and (iii) evaluate its overlap with wrong-class categories. If the predicted overlap exceeds a threshold θ, we create a new category instead of absorbing $\boldsymbol{x}'$ into the existing one. Overlap is computed as the normalized L^1 intersection of hyperbox bounds, directly penalizing expansions that intrude into wrong-class regions. This rule preserves class separation while still incorporating informative adversarial samples.

Table II summarizes the high-ϵ effect on USPS. At ϵ=0.30, separation-aware training with θ=0.01 achieves 11.0±1.4% robust accuracy, compared to 5.4±0.6% for two-stage selective training and 2.6±0.9% for standard selective training, indicating that high-ϵ robustness can benefit from explicit geometric constraints when overlap becomes a dominant failure mode.

Match score inversion and implications for rejection. Match-score statistics reveal a caveat for rejection-based defenses. In vanilla models, adversarial examples typically yield lower match scores than clean samples (USPS: 0.987 vs.

TABLE II: USPS High-ϵ Robustness (mean±std over 3 seeds).

Method	Cats	@.25	@.30
Selective (on)	11.5K	4.8±0.8	2.6±0.9
Two-Stage Sel.	13.7K	11.7±1.0	5.4±0.6
Sep-Aware	12.6K	**11.9±0.4**	**11.0±1.4**

@.25/@.30 = adversarial accuracy (%) at $\epsilon = 0.25/0.30$. Sep-Aware uses $\theta = 0.01$ and checks overlap before absorbing adversarial examples.

0.997), suggesting that thresholding match may reject many adversarial inputs. After selective adversarial training, however, this relation can invert: adversarial examples can attain higher match scores than clean samples (USPS: 0.983 vs. 0.968), consistent with absorption of adversarial patterns into existing categories. Consequently, a match-threshold detector can fail badly: AUC drops from 0.72 on vanilla to 0.38 after selective training. Rejection thresholds therefore must be calibrated for the specific trained variant rather than transferred from a vanilla model.

V. Experimental Setup

We evaluate on three benchmarks, namely USPS (7,291 train / 2,007 test, 16×16), MNIST (60,000 / 10,000, 28×28), and Fashion-MNIST (60,000 / 10,000, 28×28). All images are normalized to $[0, 1]$ and flattened. We use Fuzzy ARTMAP with $\alpha = 10^{-3}$, $\beta = 1.0$, $\rho_{ab} = 1.0$. We sweep vigilance $\rho_a \in \{0.0, 0.1, \ldots, 0.9\}$ for vanilla models; baseline vulnerability results (Figure 1) show $\rho_a = 0.5$–0.9 for clarity, as lower vigilance yields poor clean accuracy. Defense methods are evaluated at $\rho_a = 0.9$, which achieves the best robustness. For white-box attacks, we use WB-Softmax with temperature $\tau = 0.01$ and $\epsilon \in \{0.05, \ldots, 0.35\}$. For black-box, we use PGD transfer from SimpleCNN and LRS-enhanced surrogates.

We compare six defense methods (Table III). Vanilla provides the no-defense baseline. Adversarial training (AdvTrain) trains on all adversarial examples, evaluated in offline and online variants. Selective training filters to adversarial examples that fool the model. Two-stage selective training uses progressive epsilon scheduling ($\epsilon \in [0.05, 0.15]$ then $[0.15, 0.35]$) with selective filtering. Separation-aware training checks overlap before absorbing adversarial examples. For online variants, we generate WB-Softmax adversarial examples on-the-fly and discard them immediately, maintaining single-pass streaming.

All experiments use three training seeds (42, 123, 456) with a fixed evaluation subset of n=1000 clean-correct test samples, chosen to keep PGD-20 evaluation tractable on models with up to 121K categories. Robust accuracy is computed on the subset of test samples correctly classified at $\epsilon = 0$ (clean-correct subset). This conditional metric ensures consistent comparisons across methods with different clean accuracies, but note that cross-ρ comparisons reflect conditional robustness given correct clean prediction.

We evaluate white-box robustness under WB-Softmax PGD, our strongest differentiable attack (89–100% success on vanilla models at $\epsilon = 0.35$). Black-box robustness is evaluated primarily under CNN-transfer PGD and LRS-transfer PGD; Square attack is used as an additional black-box baseline for vanilla models (Table V). All adversarial training uses WB-Softmax attacks to ensure defenses are evaluated against the strongest available white-box attacker.

VI. Results and Discussion

This section establishes baseline vulnerability of vanilla Fuzzy ARTMAP, compares defenses under WB-Softmax, and discusses category dynamics and anti-gradient-masking verification.

A. Baseline Vulnerability

Figure 1 shows vanilla Fuzzy ARTMAP accuracy versus ϵ across vigilance levels under both WB-Softmax and black-box transfer attacks. WB-Softmax PGD achieves attack success rates exceeding black-box transfer across all datasets, confirming that gradients provide useful signal. Higher vigilance improves robustness under both attack types.

B. Defense Comparison Under WB-Softmax

Table III compares all methods at $\rho = 0.9$ under WB-Softmax attack, and Figure 2 shows accuracy-vs-ϵ curves for key methods. We report clean accuracy, adversarial accuracy at ϵ=0.30, AURAC, and category count. Separation-aware training is evaluated as a USPS case study (Table II).

For clarity, Table III reports robustness under the adaptive white-box WB-Softmax threat model, whereas Table IV reports transfer-based black-box robustness. Table IV evaluates the same defenses under black-box transfer attacks. The robustness rankings under the black-box transfer suite are notably different from white-box evaluation. Offline adversarial training collapses under WB-Softmax, yet achieves the highest AURAC on USPS (63.8%) and Fashion-MNIST (82.1%) under transfer attacks. This reversal suggests a mismatch in offline training: adversarial examples are generated against a fixed clean model, but the decision boundary shifts after incremental updates, so the resulting robustness may not transfer to adaptive white-box attacks on the *final* model. Transfer-based black-box attacks are more model-agnostic, so offline training can still help by learning general perturbation patterns. This further supports the use of online training, which generates adversarial examples against the *current* model state and better matches the white-box threat model.

Across all three datasets, offline adversarial training degrades robustness despite roughly doubling the category count (Table III). This indicates that perturbations generated against a fixed clean model are poorly aligned with the final incrementally updated ARTMAP model; the additional categories appear to memorize stale perturbations rather than improve robustness against adaptive attacks.

Online adversarial training avoids this fixed-model mismatch but remains dataset-dependent, improving USPS and Fashion-MNIST while degrading MNIST. This suggests that robustness depends not only on generating current-model perturbations, but also on which adversarial samples are admitted and when they are introduced.

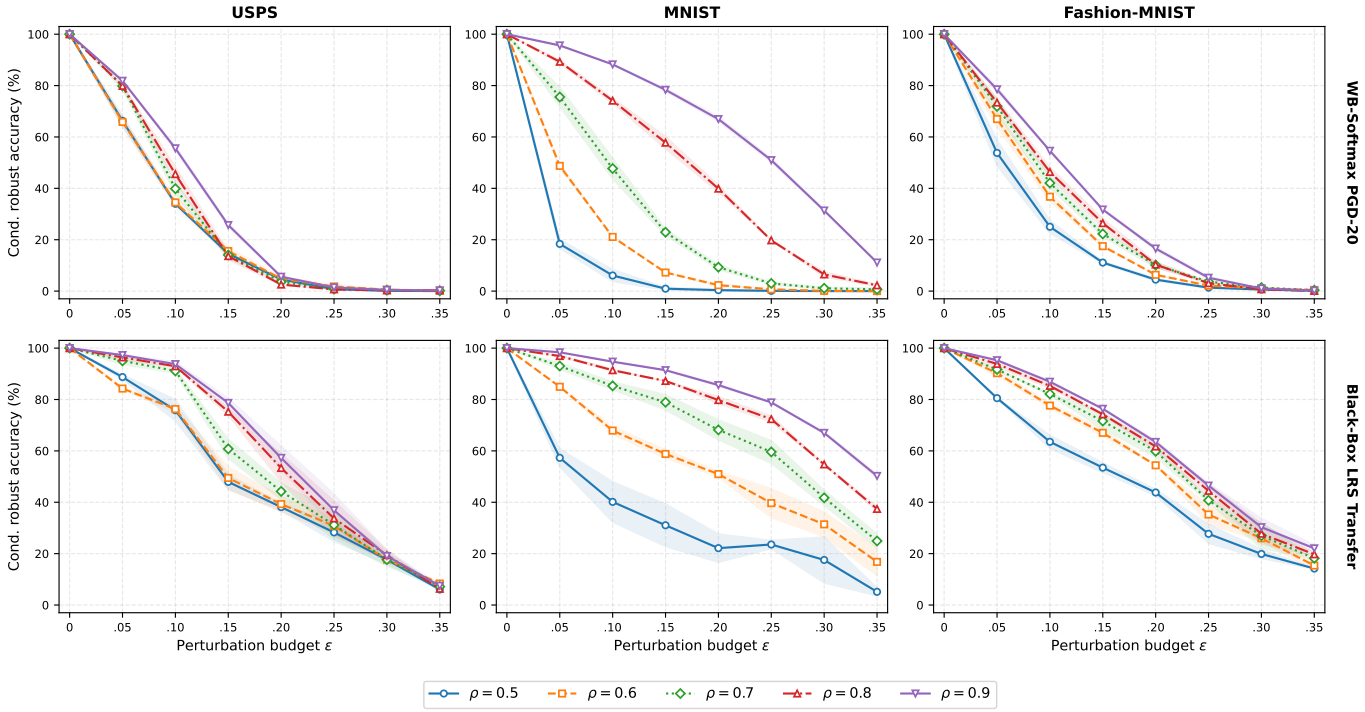

Fig. 1: Baseline vulnerability of vanilla Fuzzy ARTMAP ($\epsilon \leq 0.35$). **Top**: White-box Softmax PGD (WB-Softmax) attack with temperature $\tau = 0.01$. **Bottom**: Black-box transfer PGD attack.

TABLE III: Defense Comparison under WB-Softmax PGD-20 ($\rho = 0.9$). Mean±std over 3 seeds

	USPS				MNIST				Fashion-MNIST			
Method	Clean	@.30	AURAC	Cats	Clean	@.30	AURAC	Cats	Clean	@.30	AURAC	Cats
Vanilla	92.3±0.6	0.5±0.1	21.7±0.2	6K	94.9±0.2	31.7±0.6	61.5±0.1	56K	80.7±0.5	0.9±0.2	24.4±0.6	57K
AdvTrain (off)	92.4±0.1	0.3±0.1	12.6±0.4	12K	94.5±0.1	0.0±0.0	14.8±0.2	103K	82.2±0.2	1.1±0.2	14.9±0.2	109K
AdvTrain (on)	92.9±0.4	10.6±1.2	25.9±0.6	13K	94.1±0.1	21.1±4.4	51.2±0.6	111K	82.8±1.0	20.7±0.5	34.0±0.2	111K
Selective (off)	92.2±0.0	3.0±0.9	25.4±0.7	11K	94.5±0.0	26.3±0.6	53.6±0.1	94K	82.4±0.1	15.3±1.2	35.1±0.5	101K
Selective (on)	92.3±0.8	2.6±0.9	24.7±0.4	12K	94.9±0.2	26.7±1.1	53.0±0.6	95K	81.1±0.4	17.0±0.8	34.8±0.1	101K
Sep-Aware	92.5±0.4	**11.0±1.4**	26.7±0.5	13K	94.2±0.1	25.1±2.3	51.9±0.2	111K	82.9±0.9	20.5±1.0	34.7±0.5	111K
Two-Stage Sel.	92.1±0.1	5.4±0.6	**28.2±0.2**	14K	94.5±0.1	**45.8±1.6**	**64.5±0.7**	89K	82.4±0.1	**23.0±1.0**	**41.3±0.2**	121K

@.30 = adversarial accuracy (%) at ϵ=0.30. AURAC (%) = Area Under Robust Accuracy Curve ($\epsilon \in [0.05, 0.35]$). Cats = category count (K = thousands). Bold = best per dataset.

Two-stage selective training achieves the highest AURAC on all three datasets (Table III). On USPS it achieves 28.2±0.2% (+30% over vanilla), on MNIST 64.5±0.7% (+5% over vanilla, +26% over online AdvTrain), and on Fashion-MNIST 41.3±0.2% (+69% over vanilla). The progressive epsilon scheduling allows the model to establish moderate robustness before facing strong attacks, while selective filtering ensures training focuses on boundary regions where the model is vulnerable. Empirically, the two-stage schedule appears to reduce the instability observed when training is exposed immediately to high-ϵ attacks, while selective filtering prevents over-updating on adversarial samples that do not change the predicted label.

While two-stage training optimizes overall robustness (AURAC), separation-aware training targets high-ϵ performance by preventing category overlap during adversarial absorption. On USPS, cluster validity analysis reveals that selective training can induce separation collapse (Table I). Separation-aware training achieves 11.0±1.4% at ϵ=0.30, substantially outperforming two-stage selective (5.4±0.6%) and standard selective (2.6±0.9%) at this perturbation level (Table II). On MNIST and Fashion-MNIST, separation-aware training achieves AURAC of 51.9% and 34.7% respectively, comparable to selective training but without exceeding it—suggesting that separation constraints provide targeted high-ϵ benefits rather than broad robustness gains in high-category regimes. This demonstrates how ART's interpretable geometry enables targeted mitigations for specific operating points.

TABLE IV: Defense Comparison under Black-Box Transfer PGD-20 ($\rho = 0.9$). Mean±std over 3 seeds

	USPS				MNIST				Fashion-MNIST			
Method	Clean	@.30	AURAC	Cats	Clean	@.30	AURAC	Cats	Clean	@.30	AURAC	Cats
Vanilla	92.2±0.1	9.1±0.3	48.4±0.2	6K	94.5±0.1	85.8±0.4	92.1±0.2	56K	82.2±0.2	67.8±0.5	79.8±0.3	57K
AdvTrain (off)	92.4±0.2	**41.5±1.9**	**63.8±0.8**	12K	94.5±0.1	81.1±0.6	89.2±0.3	103K	82.2±0.3	**72.5±0.2**	**82.1±0.1**	109K
AdvTrain (on)	92.9±0.5	25.7±1.1	57.3±0.6	13K	94.1±0.2	79.4±1.6	88.2±1.0	111K	82.8±1.2	68.6±1.6	79.9±0.9	111K
Selective (off)	92.1±0.1	29.2±0.4	58.3±0.2	11K	94.5±0.0	88.7±1.2	93.3±0.2	94K	82.4±0.1	67.5±0.2	79.9±0.2	101K
Selective (on)	92.5±0.2	24.0±0.5	56.3±0.3	12K	94.5±0.1	89.2±0.3	93.5±0.2	95K	81.7±0.2	68.5±0.4	80.4±0.2	101K
Sep-Aware	92.8±0.5	25.8±0.3	57.8±0.3	13K	94.2±0.2	78.0±1.9	87.5±0.7	111K	82.9±1.2	68.3±2.2	79.7±1.1	111K
Two-Stage Sel.	92.1±0.1	30.0±0.7	59.4±0.2	14K	94.5±0.0	**89.4±0.4**	**93.7±0.2**	89K	82.4±0.2	69.0±0.4	80.5±0.2	121K

Transfer attacks use SimpleCNN surrogate trained on each dataset. @.30 = adversarial accuracy (%) at ϵ=0.30. AURAC (%) = Area Under Robust Accuracy Curve ($\epsilon \in [0.05, 0.35]$). Cats = category count (K = thousands). Bold = best per dataset.

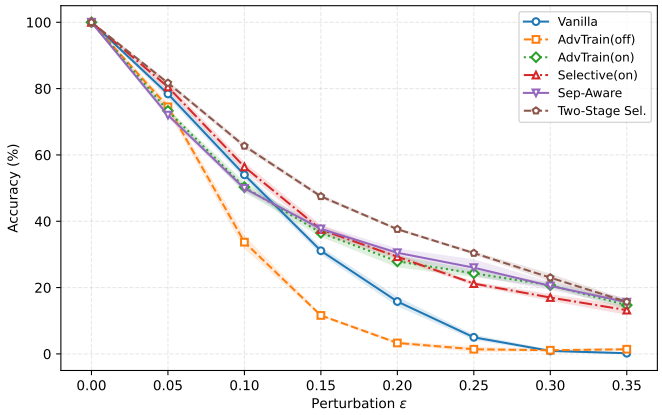

Fig. 2: Defense comparison on Fashion-MNIST under WB-Softmax PGD-20 (ρ=0.9). Two-stage selective training achieves the highest robust accuracy, while offline adversarial training collapses below vanilla. Mean over 3 seeds. Due to space constraints, we plot accuracy–ϵ curves only for Fashion-MNIST; full tabular results for all datasets are reported in Tables III–IV.

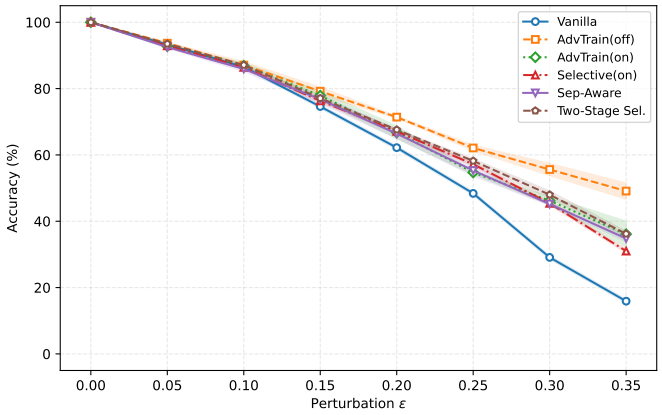

Fig. 3: Defense comparison on Fashion-MNIST under LRS Transfer PGD-20 (ρ=0.9). Offline adversarial training achieves highest accuracy under black-box transfer attacks, reversing its poor white-box performance. Mean over 3 seeds.

C. Additional Analyses and Verification

Table V compares attack methods on vanilla models at $\rho = 0.9$. WB-Softmax PGD achieves 89–100% attack success across datasets, exceeding transfer and query-based black-box baselines, with transfer close on USPS (97-98% vs. 100% at ϵ=0.35). This confirms that effective white-box attacks exist when the objective properly aligns with ART's decision rule.

TABLE V: Attack Strength Comparison (Vanilla Models, $\rho = 0.9$, $\epsilon = 0.35$)

Attack	USPS	MNIST	FMNIST
WB-Softmax PGD	100	89	100
BB Transfer PGD	97	18	41
LRS Transfer PGD	98	19	84
Square (query)	64	28	61

Attack success rate (%) on n=1000 clean-correct samples. PGD uses 20 steps with step size $\eta = \epsilon/4$. Square Attack uses 5000 queries/sample.

All adversarial training variants increase category counts (Table III). The critical observation is that category growth alone does not predict robustness. Offline AdvTrain creates the most categories on MNIST (103K) yet achieves the worst AURAC (14.8%). Two-stage selective training creates moderate category growth (1.6–2.2×) while achieving the best robustness, suggesting that *how* categories are created matters more than how many.

Table VI verifies genuine robustness through adapted sanity checks [5]. All four criteria pass on all three datasets: WB-Softmax PGD outperforms FGSM, WB-Softmax matches or exceeds black-box transfer, smooth surrogates outperform hard operations, and accuracy degrades smoothly with ϵ.

TABLE VI: Anti-Gradient-Masking Sanity Checks (ρ=0.9)

Sanity Check	USPS	MNIST	FMNIST
PGD $\geq$ FGSM	✓	✓	✓
WB $\geq$ BB transfer	✓	✓	✓
Smooth $>$ Hard ops	✓	✓	✓
Smooth acc-vs-ϵ	✓	✓	✓

Evaluated on vanilla models at ϵ=0.35. WB = WB-Softmax, BB = black-box.

VII. Conclusion

We presented a systematic study of adversarial robustness in Fuzzy ARTMAP under a true incremental learning setting. WB-Softmax PGD provides an adaptive white-box attack by relaxing ART's winner-take-all competition into a differentiable class-level objective, achieving 89–100% attack success on vanilla models.

Our investigation revealed three key findings. First, training protocol is critical. Offline adversarial training universally collapses robustness across all datasets despite substantial category growth, while online training yields inconsistent, dataset-dependent results. Second, progressive training improves consistency. Two-stage selective training, which gradually increases attack strength while filtering to successful attacks, achieves the best overall robustness (AURAC) across all three datasets. Third, ART's explicit category geometry enables targeted diagnosis and mitigation; geometry monitoring reveals separation collapse, and separation-aware training improves high-ϵ robustness on USPS.

Overall, our results highlight that interpretable, prototype-based incremental learners exhibit distinct robustness behaviors from batch-trained deep networks. The WB-Softmax attack and the training analyses in this work provide a foundation for developing and evaluating adversarial defenses tailored to ART and other non-differentiable, prototype-based architectures operating in streaming environments. Future work will extend these results to larger-scale datasets, additional modalities, and other ART/ARTMAP variants.

Acknowledgment

This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-22-2-0209. This research was supported by NSF grant 2420248 and by the Kummer Institute, Mary Finley Endowment, and Intelligent Systems Center of the Missouri University of Science and Technology.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

The computation for this work was performed on the high-performance computing infrastructure provided by Research Support Solutions at Missouri University of Science and Technology https://doi.org/10.71674/PH64-N397

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018.

[3] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 39–57.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[5] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 274–283.

[6] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "RobustBench: a standardized adversarial robustness benchmark," in *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[7] M. Zhao, L. Zhang, J. Ye, H. Lu, B. Yin, and X. Wang, "Adversarial training: A survey," *arXiv preprint arXiv:2410.15042*, 2024.

[8] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

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

[10] X. Mi, F. Tang, Z. Yang, D. Wang, J. Cao, P. Li, and Y. Liu, "Adversarial robust memory-based continual learner," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[11] J. Bang, H. Koh, S. Park, H. Song, J.-W. Ha, and J. Choi, "Online continual learning on a contaminated data stream with blurry task boundaries," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9265–9274.

[12] S. Grossberg, "Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world," *Neural Networks*, vol. 37, pp. 1–47, 2013.

[13] L. E. Brito da Silva, I. Elnabarawy, and D. C. Wunsch II, "A survey of adaptive resonance theory neural network models for engineering applications," *Neural Networks*, vol. 120, pp. 167–203, 12 2019.

[14] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 698–713, 1992.

[15] N. M. Melton, D. Tanksley, and D. C. Wunsch II, "Adaptive resonance lib: A Python package for adaptive resonance theory (ART) models," *Journal of Open Source Software*, vol. 10, no. 114, p. 7764, 2025.

[16] N. M. Melton, L. E. Brito da Silva, S. Petrenko, and D. C. Wunsch II, "Deep ARTMAP: Generalized hierarchical learning with adaptive resonance theory," *arXiv preprint arXiv:2503.07641*, 2025.

[17] S. Petrenko, L. E. Brito da Silva, and D. C. Wunsch II, "DeepART: Deep gradient-free local learning with adaptive resonance," *Neural Networks*, vol. 190, p. 107580, 2025.

[18] N. M. Melton, L. E. Brito da Silva, and D. C. Wunsch II, "An extensive analysis of match-tracking methods for ARTMAP," in *2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM)*. IEEE, 2025, pp. 1–8.

[19] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Networks*, vol. 4, no. 6, pp. 759–771, 1991.

[20] T. Wu, T. Luo, and D. C. Wunsch II, "LRS: Enhancing adversarial transferability through Lipschitz regularized surrogate," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 6, 2024, pp. 6135–6143.

[21] 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," *arXiv preprint arXiv:2310.17626*, 2024.

[22] L. E. Brito da Silva, N. M. Melton, and D. C. Wunsch II, "Incremental cluster validity indices for online learning of hard partitions: Extensions and comparative study," *IEEE Access*, vol. 8, pp. 22 025–22 047, 2020.

[23] L. E. Brito da Silva, N. Rayapati, and D. C. Wunsch II, "iCVI-ARTMAP: Using incremental cluster validity indices and adaptive resonance theory reset mechanism to accelerate validation and achieve multiprototype unsupervised representations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 9757–9770, 2023.