ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11463281

PAST AS PRIOR: REWEIGHTED PROXY GUIDANCE FOR STABLE ADVERSARIAL TRAINING

Yaohua Liu[⋆] *Jiaxin Gao*[†*]

[⋆] School of Computing and Data Science, The University of Hong Kong
[†] Department of Data Science & AI, The Hong Kong Polytechnic University

ABSTRACT

Adversarial robustness is critical for the reliable deployment of deep neural networks in safety-sensitive applications, with adversarial training (AT) being the dominant defense technique. However, existing AT methods still suffers from unstable convergence, large variance, and catastrophic overfitting. To alleviate these limitations, we propose Reweighted Proxy Guidance (RPG), which treats the immediately preceding model as a history-driven prior to steer updates toward more robust solutions. At its core, a Reweighted Differential Unit (RDU) forms a reweighted differential between the current parameters and a proxy-induced response, providing a flexible update rule compatible with both single-step and multi-step AT. We further introduce a teacher-free self-distillation defense objective aligned with the proxy to regularize the learning trajectory and mitigate catastrophic overfitting. Extensive evaluations showcase RPG consistently improves performance and stabilizes training across diverse datasets, backbones, and attack budgets (e.g., 20.3% enhancement in robust accuracy on CIFAR100 dataset over PGD-AT).

***Index Terms*—** Adversarial training, classification, self-distillation, reweighted differential unit

1. INTRODUCTION

Deep neural networks in safety–critical settings are vulnerable to small adversarial perturbations, degrading downstream performance. Reliable defenses [1] are therefore indispensable for safe deployment. Among them, Adversarial Training (AT) [2, 3] is the prevailing paradigm, yet practical deployment remains hampered by instability and efficiency constraints. We next review related work to position our focus.

1.1. Related Work

Adversarial attacks. In this paper we primarily consider white-box, gradient-based attacks used for training and evaluation. FGSM [2] applies a single signed–gradient step; PGD [4] performs multi–step projected updates (often with random starts) and serves as the inner maximization for AT; AutoAttack [5] aggregates strong, parameter–free components to provide a reliable benchmark.

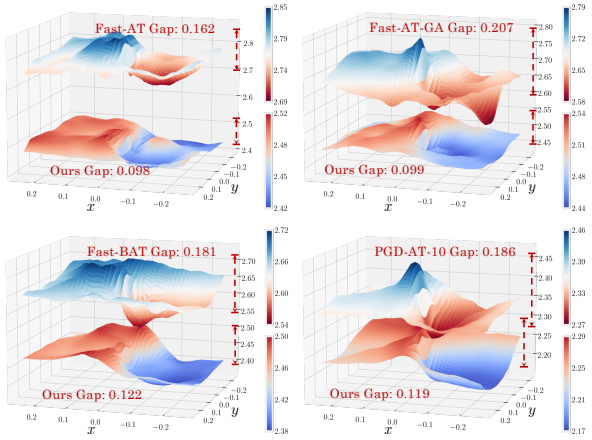

Fig. 1. Adversarial loss landscapes of 4 representative adversarial training baselines and our RPG-augmented variants on CIFAR-10 (ϵ=8/255) dataset. RPG yields lower, smoother surfaces and consistently smaller mini-max gaps over $(x, y) \in [-0.25, 0.25]$, indicating improved robustness.

Adversarial defenses. Single-step AT [6, 7] is compute-efficient but prone to catastrophic overfitting [8]. Strengthening efforts include gradient-alignment regularization [9] and bi-level/implicit-gradient updates [10]. Multi-step PGD-based training [4, 11] attains higher robustness at higher cost. Other lines reshape the objective or supervision—TRADES for robustness–accuracy balance [12], misclassification-aware training [13], and teacher-based distillation [14]. Recent work revisits robustness distillation with fairness [15], tunes weight decay [16], and injects prior knowledge for Fast-AT [17]. Historical-state approaches [18] average checkpoints (SWA [19]) or add retrospective losses [20]. Despite progress, AT remains unstable with high seed variance; single-step often exhibits catastrophic overfitting, and multi-step, though stronger, still suffers stability issues.

1.2. Contributions

We address training instability and catastrophic overfitting by introducing Reweighted Proxy Guidance (RPG). In Fig. 1,

* Corresponding author.

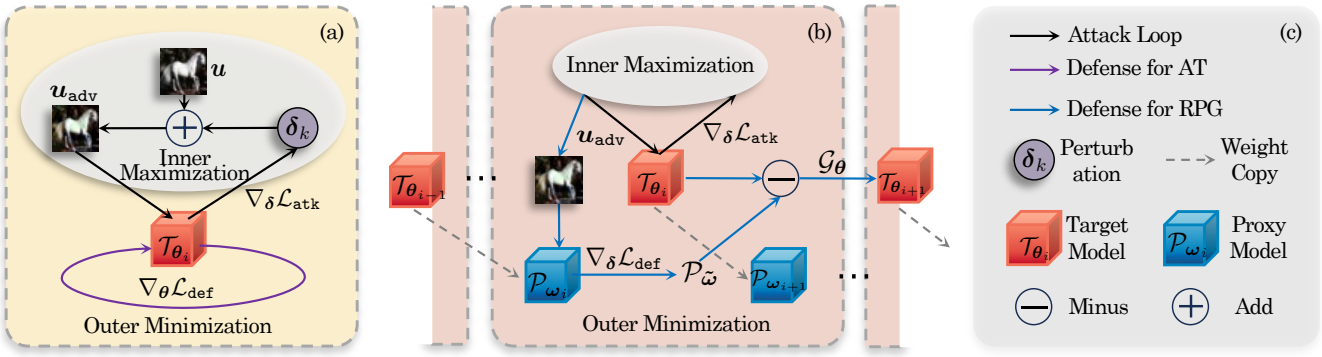

Fig. 2. Pipelines. (a) Standard AT: Minimization and maximization with the target model; (b) RPG: Outer minimization guided by the reweighted differential unit via a proxy from the previous state; (c) Notation of symbols.

we visualize the adversarial loss along the sign-gradient direction and an orthogonal random direction, i.e., $\mathcal{L}_{\texttt{atk}}(\boldsymbol{u} + x\,\vec{\boldsymbol{\iota}} + y\,\vec{\boldsymbol{o}})$ with $\vec{\boldsymbol{\iota}} = \operatorname{sgn}(\nabla_{\mathbf{I}}\mathcal{L}_{\texttt{atk}}(\boldsymbol{u}))$ and $\vec{\boldsymbol{o}} \sim \text{Rademacher}$. RPG-augmented models yields lower, smoother adversarial loss surfaces. We summarize our contributions as follows.

1. We present the RPG framework that treats the historical model state as a defense prior, providing a new perspective to enhance robustness and stabilize optimization.

2. We develop the reweighted differential unit, a flexible update rule that forms the differential between the current parameters and a proxy-induced response, consistently improving single- and multi-step AT methods.

3. We design a teacher-free self-distillation objective aligned with the proxy to regularize the optimization trajectory, mitigating catastrophic overfitting and further stabilizes training with minor overhead.

4. Across diverse datasets, backbones, and attack strengths, RPG consistently improves robust accuracy with stabilized convergence behavior; under AutoAttack with $\epsilon = 16/255$, performance gains reach 9.2% on CIFAR-10 and 20.3% on CIFAR-100 over PGD-AT.

2. REWEIGHTED PROXY GUIDANCE

2.1. Motivation: Past as Prior

We adopt the standard min–max formulation of AT [4]:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{u}_i,\boldsymbol{v}_i)}\Big[\max_{\boldsymbol{\delta}\in\mathcal{S}} \mathcal{L}\big(\mathcal{T}_{\boldsymbol{\theta}}(\boldsymbol{u}_i{+}\boldsymbol{\delta}), \boldsymbol{v}_i\big)\Big], \tag{1}$$

where $\boldsymbol{u}_i$ is the input with ground-truth label $\boldsymbol{v}_i$, $\mathcal{T}_{\boldsymbol{\theta}}$ denotes the target model with parameters $\boldsymbol{\theta}$, and $\boldsymbol{\delta}$ is the perturbation within the constraint set $\mathcal{S}{=}\{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_\rho \le \epsilon\}$. Then the inner maximization is typically approximated by K-step PGD as:

$$\boldsymbol{\delta}_{k+1} \leftarrow \Pi_\epsilon\Big(\boldsymbol{\delta}_k + \alpha\,\operatorname{sgn}\nabla_{\boldsymbol{\delta}}\mathcal{L}\big(\mathcal{T}_{\boldsymbol{\theta}}(\boldsymbol{u}_i{+}\boldsymbol{\delta}_k), \boldsymbol{v}_i\big)\Big). \tag{2}$$

Past as prior. Attacks in Eq. (2) are tailored to the *current* state $\boldsymbol{\theta}_i$, while the immediately preceding state is inaccessible to the inner loop. We denote this predecessor as the proxy $\mathcal{P}_{\boldsymbol{\omega}_i}$ with $\boldsymbol{\omega}_i{=}\boldsymbol{\theta}_{i-1}$. Empirically, adversarial examples optimized on $\boldsymbol{\theta}_i$ transfer poorly to black-box models, suggesting the proxy provides a defensive prior that is less aligned with the current attack.

Lemma 1. We assume $g_i^{\mathrm{T}}(\boldsymbol{x}){=}\nabla_{\boldsymbol{x}}\mathcal{L}(\mathcal{T}_{\boldsymbol{\theta}_i}(\boldsymbol{x}), \boldsymbol{v}_i)$ and $g_i^{\mathrm{P}}(\boldsymbol{x}){=}\nabla_{\boldsymbol{x}}\mathcal{L}(\mathcal{P}_{\boldsymbol{\omega}_i}(\boldsymbol{x}), \boldsymbol{v}_i)$. For ℓ_∞ step $\boldsymbol{\delta}_i = \epsilon\,\operatorname{sgn}\big(g_i^{\mathrm{T}}(\boldsymbol{u}_i)\big)$,

$$\langle g_i^{\mathrm{P}}(\boldsymbol{u}_i),\, \boldsymbol{\delta}_i\rangle \le \epsilon\,\|g_i^{\mathrm{P}}(\boldsymbol{u}_i)\|_1, \tag{3}$$

with equality only when $g_i^{\mathrm{P}}(\boldsymbol{u}_i)$ and $g_i^{\mathrm{T}}(\boldsymbol{u}_i)$ have identical sign patterns. *Implication.* Unless gradients are perfectly sign-aligned, the target–tailored step is strictly less effective on the proxy, revealing a history-driven prior. Therefore, as shown in Fig. 2, we further modify the inner maximization via a proxy-guided reweighted update to enhance robustness.

2.2. Reweighted Differential Unit (RDU)

Motivated by the misalignment observation in Lemma 1, we set the proxy as the previous target state: $\boldsymbol{\omega}_i{=}\boldsymbol{\theta}_{i-1}$ with $\boldsymbol{\omega}_0{=}\boldsymbol{\theta}_0$. Let $\mathcal{L}_{\texttt{atk}}$ and $\mathcal{L}_{\texttt{def}}$ denote attack and defense losses (same form $\mathcal{L}$ in practice). After generating $\boldsymbol{\delta}_K$ via the standard inner maximization, we take a single descent step on the proxy to obtain fast weights

$$\tilde{\boldsymbol{\omega}} = \boldsymbol{\omega}_i - \beta\,\nabla_{\boldsymbol{\omega}}\,\mathcal{L}_{\texttt{def}}\big(\mathcal{P}_{\boldsymbol{\omega}_i}(\boldsymbol{u}_i{+}\boldsymbol{\delta}_K),\, \boldsymbol{v}_i\big). \tag{4}$$

Then we define RDU as $\mathcal{G}_{\boldsymbol{\theta}} = \boldsymbol{\theta}_i{-}\tilde{\boldsymbol{\omega}}$, and update sequentially

$$\boldsymbol{\omega}_{i+1} = \boldsymbol{\theta}_i,\ \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \boldsymbol{\gamma}\,\mathcal{G}_{\boldsymbol{\theta}}. \tag{5}$$

As shown in Alg. 1, the inner maximization remains unchanged, while the outer update is *history-guided* by the proxy; this steers the step away from the current attack–aligned direction and improves stability of convergence behavior.

Self-Distillation Regularization. To mitigate proxy drift on adversarial inputs, we introduce a teacher-free self-distillation

Table 1. We report SA and RA for PGD-AT and its RPG-augmented variant on CIFAR-10/100, and the Average RA (%).

CIFAR-10 dataset, PARN-18 trained with $\epsilon = 8/255$								
Method	SA (%)	PGD-10 (%)		PGD-50 (%)		AutoAttack (%)		Avg. (%)
		$\epsilon = 8/255$	$\epsilon = 16/255$	$\epsilon = 8/255$	$\epsilon = 16/255$	$\epsilon = 8/255$	$\epsilon = 16/255$	
PGD-AT	81.95±0.74	51.92±0.30	20.31±0.75	50.76±0.33	15.68±0.41	47.09±0.54	13.09±0.43	33.14
Ours	**82.19**±0.90	**53.23**±0.20	**22.20**±0.37	**52.14**±0.10	**17.59**±0.57	**47.71**±0.22	**14.30**±0.03	**34.53**
CIFAR-100 dataset, PARN-18 trained with $\epsilon = 8/255$								
Method	SA (%)	PGD-10 (%)		PGD-50 (%)		AutoAttack (%)		Avg. (%)
		$\epsilon = 8/255$	$\epsilon = 16/255$	$\epsilon = 8/255$	$\epsilon = 16/255$	$\epsilon = 8/255$	$\epsilon = 16/255$	
PGD-AT	**49.46**±0.48	25.84±0.43	9.98±0.43	25.38±0.39	8.75±0.38	21.11±0.05	6.67±0.23	16.29
Ours	48.15±0.42	**31.27**±0.53	**14.90**±0.26	**30.86**±0.62	**13.57**±0.52	**23.11**±0.18	**8.03**±0.40	**20.29**

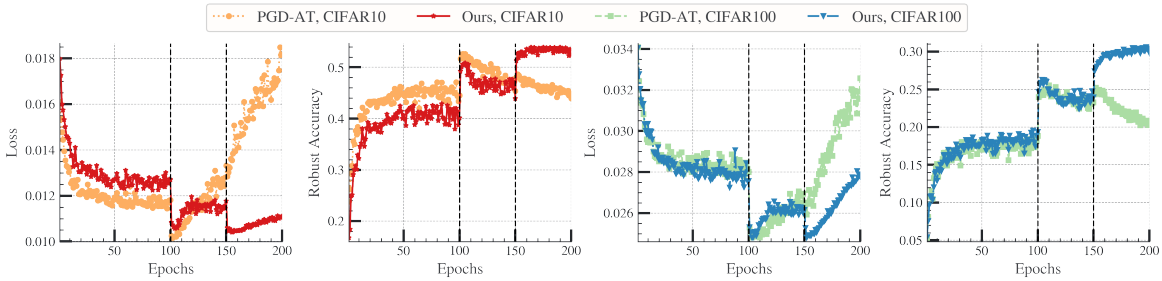

Fig. 3. Comparison of the convergence behavior of test loss and RA , $\epsilon = 8/255$ on CIFAR10 dataset and CIFAR100 dataset. The black dashed line denotes the epoch where multi-step learning rate decays.

Algorithm 1 Reweighted Proxy Guidance Framework

Require: epochs $\mathcal{J}$; batches $\{(\boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i=1}^{\mathcal{M}}$; attack steps K; step sizes $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$; target $\mathcal{T}_{\boldsymbol{\theta}}$, proxy $\mathcal{P}_{\boldsymbol{\omega}}$

Initialize $\boldsymbol{\theta}_0$; set $\boldsymbol{\omega}_0 = \boldsymbol{\theta}_0$
for $j = 0$ **to** $\mathcal{J} - 1$ **do**
  for $i = 0$ **to** $\mathcal{M} - 1$ **do**
    Initialize $\boldsymbol{\delta}_0 \sim \mathrm{Unif}(-\boldsymbol{\epsilon}, \boldsymbol{\epsilon})$
    for $k = 0$ **to** $K - 1$ **do**
      $\boldsymbol{\delta}_{k+1} = \Pi_{\boldsymbol{\epsilon}}(\boldsymbol{\delta}_k + \boldsymbol{\alpha} \cdot \mathrm{sgn}(\nabla_{\boldsymbol{\delta}} \mathcal{L}_{\mathtt{atk}}(\mathcal{T}_{\boldsymbol{\theta}_i}(\boldsymbol{u}_i + \boldsymbol{\delta}_k), \boldsymbol{v}_i)))$
    end for
    $\tilde{\boldsymbol{\omega}} = \boldsymbol{\omega}_i - \boldsymbol{\beta} \nabla_{\boldsymbol{\omega}} \mathcal{L}_{\mathtt{def}}(\mathcal{P}_{\boldsymbol{\omega}_i}(\boldsymbol{u}_i + \boldsymbol{\delta}_K), \boldsymbol{v}_i)$
    $\mathcal{G}_{\boldsymbol{\theta}} = \boldsymbol{\theta}_i - \tilde{\boldsymbol{\omega}}$ (RDU)
    $\boldsymbol{\omega}_{i+1} = \boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \boldsymbol{\gamma} \mathcal{G}_{\boldsymbol{\theta}}$
  end for
end for

loss enforcing clean–adversarial consistency. With temperature $\tau > 0$ and ratio $\mu \in [0, 1)$,

$$\mathcal{L}_{\mathtt{SD}} = \mu \, \mathcal{L}_{\mathtt{KL}}\left(\tfrac{\mathcal{P}_\omega(u_{\mathtt{adv}})}{\tau} \,\|\, \tfrac{\mathcal{P}_\omega(u)}{\tau} \right) + (1-\mu) \, \mathcal{L}_{\mathtt{def}}(\mathcal{P}_\omega(u_{\mathtt{adv}}), v) \,. \tag{6}$$

In practice, $\mathcal{L}_{\mathtt{def}}$ in Eq. (4) is replaced by $\mathcal{L}_{\mathtt{SD}}$, which effectively alleviates overfitting while preserving task supervision.

Stability Discussion. If $\mathcal{L}_{\mathtt{def}}$ is L-smooth in $\boldsymbol{\omega}$,

$$\left\| \nabla_{\boldsymbol{\omega}} \mathcal{L}_{\mathtt{def}}(\boldsymbol{\omega}_i) - \nabla_{\boldsymbol{\omega}} \mathcal{L}_{\mathtt{def}}(\boldsymbol{\omega}_{i-1}) \right\| \le L \left\| \boldsymbol{\omega}_i - \boldsymbol{\omega}_{i-1} \right\|. \tag{7}$$

From Eq. (4), $\|\tilde{\boldsymbol{\omega}}_{i+1} - \tilde{\boldsymbol{\omega}}_i\| \le \beta L \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_{i-1}\|$, so $\{\tilde{\boldsymbol{\omega}}_i\}$ is bounded whenever the proxy drift is bounded. The RPG update $\boldsymbol{\theta}_{i+1} = (1-\gamma)\boldsymbol{\theta}_i + \gamma \tilde{\boldsymbol{\omega}}_i$ is a convex combination

Table 2. SA and RA results under different perturbation sizes of AutoAttack on TinyImageNet dataset.

TinyImageNet dataset, PARN-18 trained with $\epsilon = 8/255$				
Method	SA (%)	RA (%), AutoAttack		
		$\epsilon = 2/255$	$\epsilon = 4/255$	$\epsilon = 8/255$
PGD-AT	48.09	38.82	30.18	16.46
TRADES	46.68	37.84	29.85	16.76
DyART	48.38	38.46	29.69	17.52
Ours	**49.07**	**39.38**	**30.54**	**17.57**

that smooths training; viewed as a history-aware momentum step, it pulls $\boldsymbol{\theta}_i$ toward the proxy's fast response while staying responsive to new gradients, and can be regarded as a Krasnosel'skiĭ–Mann–type iteration with bounded drift that contracts toward a history-aware fixed point.

3. EXPERIMENTS

We implement single- and multi-step AT based on RPG on CIFAR-10/100 and Tiny-ImageNet with PARN-18, and WRN-34-10 for scalability. Robustness is measured under PGD-10/50 and AutoAttack at $\epsilon \in \{8/255, 16/255\}$. We report Standard Accuracy (SA)/Robustness Accuracy (RA) as mean±std over three seeds and per-iteration runtime. Baselines include Fast-AT [6], Fast-AT-GA [9], Fast-BAT [10], and multi-step PGD-AT, TRADES [12] and DyART [21]. Multi-step AT methods also combined more recent SWA technique [22]. Training uses SGD (momentum 0.9, weight decay 5×10^{-4}), standard schedules, and γ=0.8 for RPG.

Table 3. SA and RA results of 3 baselines under PGD attack and AutoAttack. We use $m\pm n$ to denote the mean SA (i.e., m) with standard deviation (i.e., n) with 3 random seeds. Bold font indicates the best defense performance.

CIFAR-10 dataset, PARN-18 trained with $\epsilon = 8/255$									
Method	SA (%)	PGD-10 (%)		PGD-50 (%)		AutoAttack (%)		Avg.(%)	Time (Sec/ Iter)
		$8/255$	$16/255$	$8/255$	$16/255$	$8/255$	$16/255$		
Fast-AT	**83.56**±0.06	47.03±0.29	13.79±0.15	44.94±0.52	8.85±0.20	41.80±0.68	7.32±0.27	27.29	5.543×10^{-2}
Ours	81.70±0.15	**47.17**±0.15	**14.48**±0.23	**45.50**±0.04	**9.89**±0.14	**42.11**±0.19	**8.13**±0.20	**27.88**	5.719×10^{-2}
Fast-AT-GA	**81.00**±0.59	48.30±0.13	16.36±0.14	46.63±0.33	11.12±0.12	43.17±0.21	9.04±0.18	29.10	1.632×10^{-1}
Ours	79.18±0.13	**48.60**±0.06	**17.52**±0.02	**47.25**±0.09	**12.63**±0.17	**43.31**±0.23	**10.22**±0.05	**29.92**	1.643×10^{-1}
Fast-BAT	**82.01**±0.04	50.42±0.36	18.29±0.18	49.07±0.39	13.31±0.16	45.51±0.44	10.98±0.19	31.26	1.644×10^{-1}
Ours	79.72±0.14	**50.65**±0.19	**19.73**±0.05	**49.66**±0.20	**15.25**±0.20	**45.54**±0.27	**12.23**±0.27	**32.18**	1.656×10^{-1}

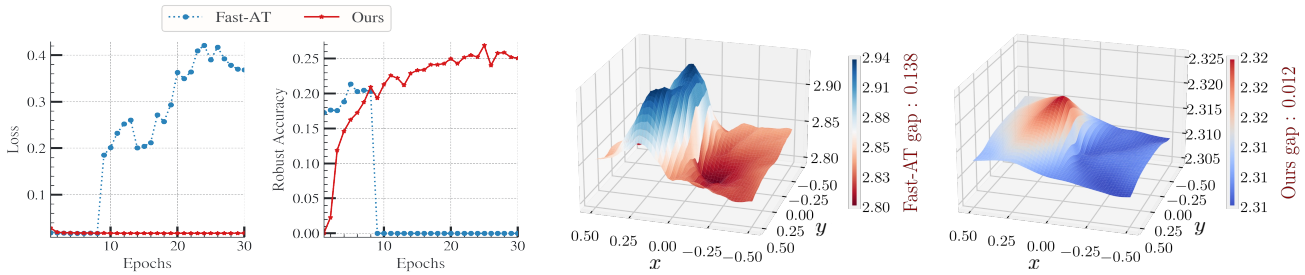

Fig. 4. Comparison of the convergence behavior and the loss landscape for Fast-AT and ours on CIFAR10 dataset under PGD-10 attack with $\epsilon = 16/255$. The loss gap is calculated within the range of $x, y \in [-0.5, 0.5]$.

Table 4. We compare single-step AT and RPG-augmented version on CIFAR10 dataset using WRN-34-10 backbone.

CIFAR-10 dataset, WRN-34-10 trained with $\epsilon = 8/255$						
Method	SA	PGD-10		PGD-50		AutoAttack
		$8/255$	$16/255$	$8/255$	$16/255$	$16/255$
Fast-AT	**80.00**	45.89	17.49	43.65	10.92	7.80
Fast-AT-GA	78.72	46.82	18.01	45.12	12.31	9.82
Fast-BAT	79.93	47.87	17.55	46.45	12.41	10.09
Ours	77.88	**49.02**	**19.23**	**47.94**	**14.15**	**11.87**

Following prior work [23], we adopt hyperparameters for self-distillation as (μ, τ)=$(0.95, 6.0)$.

3.1. Evaluation with Multi-Step AT Methods

PGD-AT with RPG. Tab. 1 shows consistent RA gains across datasets, attacks, and budgets, while SA remains comparable.
Convergence behavior. Fig. 3 indicates RPG leads to smoother loss/RA trajectories, reduced seed variance, and higher final robustness under the same schedule.
Comparison with strong baselines. On Tiny-ImageNet (Tab. 2), RPG outperforms mainstream multi-step methods with SWA against various attacks at both strengths.

3.2. Evaluation with Single-Step AT Methods

Quantitative Comparison. Tab. 3 reports consistent RA improvements for 3 mainstream baselines at both budgets, typically with lower seed variance. Moreover, the runtime overhead is negligible ($\leq$3%), confirming the efficiency of RPG.

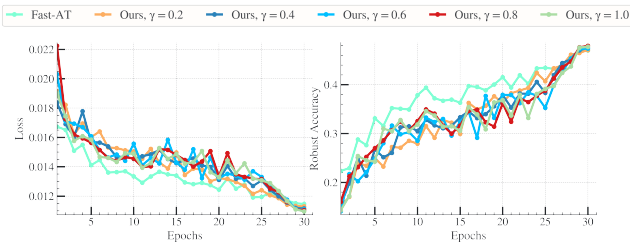

Fig. 5. Ablation of aggregation ratio γ in RPG (ϵ=$8/255$).

Alleviation of Catastrophic Overfitting. As shown in Fig. 4, at ϵ=$16/255$, injecting self-distillation regularization within RPG stabilizes Fast-AT and prevents collapse effectively.
Ablation of backbones. With WRN-34-10 backbone (Tab. 4), RPG attains the best robustness across PGD-10/50 and AutoAttack at both budgets with a small SA trade-off.
Ablation Results on γ. In Fig. 5, sweeping $\gamma \in (0, 1]$ shows RPG uniformly improves robustness and smooths convergence over Fast-AT, with the best trade-off at $\gamma \approx 0.8$.

4. CONCLUSION

We propose RPG, a general defense framework that augments AT by using the immediate predecessor as a proxy prior. An RDU steers updates and a teacher-free self-distillation term mitigates catastrophic overfitting, delivering stable optimization and consistent robustness gains with minor overhead.

5. REFERENCES

[1] Luca Bortolussi, Ginevra Carbone, Luca Laurenti, Andrea Patane, Guido Sanguinetti, and Matthew Wicker, "On the robustness of bayesian neural networks to adversarial attacks," *IEEE TNNLS*, vol. 36, no. 4, pp. 6679–6692, 2024.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[3] Qian Li, Yuxiao Hu, Yinpeng Dong, Dongxiao Zhang, and Yuntian Chen, "Focus on hiders: Exploring hidden threats for enhancing adversarial training," in *CVPR*, 2024, pp. 24442–24451.

[4] 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.

[5] Francesco Croce and Matthias Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *ICML*, 2020, pp. 2206–2216.

[6] Eric Wong, Leslie Rice, and J Zico Kolter, "Fast is better than free: Revisiting adversarial training," *arXiv preprint arXiv:2001.03994*, 2020.

[7] Hong Joo Lee, Youngjoon Yu, and Yong Man Ro, "Advancing adversarial training by injecting booster signal," *IEEE TNNLS*, vol. 35, no. 9, pp. 12665–12677, 2023.

[8] Bai Li, Shiqi Wang, Suman Jana, and Lawrence Carin, "Towards understanding fast adversarial training," *arXiv preprint arXiv:2006.03089*, 2020.

[9] Maksym Andriushchenko and Nicolas Flammarion, "Understanding and improving fast adversarial training," *NeurIPS*, vol. 33, pp. 16048–16059, 2020.

[10] Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, and Sijia Liu, "Revisiting and advancing fast adversarial training through the lens of bi-level optimization," in *ICML*. PMLR, 2022, pp. 26693–26712.

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

[12] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan, "Theoretically principled trade-off between robustness and accuracy," in *ICML*, 2019, pp. 7472–7482.

[13] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *ICLR*, 2019.

[14] Chengyu Dong, Liyuan Liu, and Jingbo Shang, "Label noise in adversarial training: A novel perspective to study robust overfitting," *NeurIPS*, vol. 35, pp. 17556–17567, 2022.

[15] Xinli Yue, Ningping Mou, Qian Wang, and Lingchen Zhao, "Revisiting adversarial robustness distillation from the perspective of robust fairness," in *NeurIPS*, 2023.

[16] Amin Ghiasi, Ali Shafahi, and Reza Ardekani, "Improving robustness with adaptive weight decay," in *NeurIPS*, 2023.

[17] Xiaojun Jia, Yong Zhang, Xingxing Wei, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao, "Improving fast adversarial training with prior-guided knowledge," *IEEE TPAMI*, vol. 46, no. 9, pp. 6367–6383, 2024.

[18] Mohammad Amin Ghiasi, Ali Shafahi, and Reza Ardekani, "Improving robustness with adaptive weight decay," *NeurIPS*, vol. 36, pp. 79067–79080, 2023.

[19] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson, "Averaging weights leads to wider optima and better generalization," *arXiv preprint arXiv:1803.05407*, 2018.

[20] Surgan Jandial, Ayush Chopra, Mausoom Sarkar, Piyush Gupta, Balaji Krishnamurthy, and Vineeth Balasubramanian, "Retrospective loss: Looking back to improve training of deep neural networks," in *ACM SIGKDD*, 2020, pp. 1123–1131.

[21] Yuancheng Xu, Yanchao Sun, Micah Goldblum, Tom Goldstein, and Furong Huang, "Exploring and exploiting decision boundary dynamics for adversarial robustness," in *ICLR*, 2023.

[22] Peng Wang, Li Shen, Zerui Tao, Shuaida He, and Dacheng Tao, "Generalization analysis of stochastic weight averaging with general sampling," in *ICML*, 2024, vol. 235, pp. 51442–51464.

[23] Biqing Qi, Bowen Zhou, Weinan Zhang, Jianxing Liu, and Ligang Wu, "Improving robustness of intent detection under adversarial attacks: A geometric constraint perspective," *IEEE TNNLS*, 2023.