Supplementary Materials of 'Mitigating Error Amplification in Fast Adversarial Training'

Mengnan Zhao[1], Lihe Zhang[2]*, Bo Wang[2], Tianhang Zheng[3]*, Hong Zhong[1], Geyong Min[4]
[1]Anhui University, Anhui, China [2]Dalian University of Technology, Liaoning, China
[3]Zhejiang University, Zhejiang, China [4]University of Exeter, UK

Unclear contribution of distribution awareness. Table 1 reports the performance comparison under various perturbation budget allocation strategies. 'Random' is a confidence-agnostic allocation that preserves the DDG budget distribution. It consistently causes catastrophic overfitting. We also include 'Percentile' (using a ratio of τ_1/B) and 'Linear' (budgets within [4/255,12/255]) schemes. 'Percentile' performs similarly to DDG, while 'Linear' improves clean accuracy but reduces robustness.

Table 1. Performance comparison under various perturbation budget allocation strategies.

Res18/CIFAR10	Clean	PGD20	PGD100	C&W	AA	CO	PGD100
Eval budgets(ξ)	-	8/255	8/255	8/255	8/255	-	12/255
DDG(Tanh)	82.7	59.6	59.3	49.9	48.1	×	43.7
Random1	70.2	47.2	47.1	42.2	41.5	✓	33.9
Random2	68.5	46.5	46.4	42.7	40.9	✓	33.9
Random3	69.6	47.7	47.6	43.1	41.7	✓	34.0
Percentile	82.8	59.6	59.4	49.4	48.1	×	43.8
Linear	84.8	57.5	57.3	49.0	47.4	×	38.6

Runtime comparison. We compare runtime with ResNet-18, trained for 110 epochs, using an Intel® Core™ Ultra 9 285K CPU and an NVIDIA RTX 5090D GPU. The runtime of DDG is comparable to that of TDAT and FGSM-RS, and is significantly more efficient than MART.

Table 2. Runtime comparison.

Time(min)	Steps	CIFAR10	CIFAR100	TinyImageNet
MART	10	120.1	120.1	142.3
FGSM-RS	1	30.3	30.3	55.0
TDAT	1	35.6	35.6	57.6
DDG	1	36.4	36.4	59.1

*corresponding author.

Performance comparison under various networks. Table 3 compares the robustness performance of different networks on the CIFAR-10 dataset. Our method outperforms the TDAT baseline across various attack settings. For ResNet-34, our approach achieves the best overall robustness, improving the accuracy under FGSM, BIM, and PGD attacks by approximately 3–4% on average, while maintaining comparable clean accuracy (83.16%). For VGG-11, similar improvements are observed, where our method yields consistent gains under most adversarial settings. Although TDAT achieves marginally higher performance under the C&W and AA attacks, our approach maintains competitive robustness. Overall, the results verify that the proposed strategy achieves a favorable balance between clean accuracy and adversarial robustness. The consistent improvements across networks further demonstrate the architecture-agnostic nature of our design.

Robustness under varied budgets. In our main paper, we employ a perturbation budget of 8/255 during training and evaluation. Additional robustness results under varying perturbation budgets on CIFAR10 and ResNet18 are reported in this appendix. Table 4 presents the quantitative evaluation. Models trained with an 8/255 budget demonstrate strong generalization across different evaluation budgets. Compared with the state-of-the-art baseline, our approach achieves substantial gains under FGSM, PGD, and BIM attacks, and delivers competitive or superior performance on C&W and PGD settings.

Ablation studies on β_1 and β_2 in Eq. (4). The main paper assigns a positive value to β_1 and a negative value to β_2, corresponding to positive reinforcement and negative suppression, respectively. Figure 1 presents the opposite setting. As shown, decreasing β_1 substantially undermines training stability, and strengthening the probability of the most probable incorrect class in high-confidence groups also leads to significant instability. In contrast, increasing the probability of the most probable incorrect class in low-confidence groups maintains the training stability but results in reduced classification accuracy. These observations collectively indicate that the supervision signal associated with the correct

Table 3. Performance comparison under various networks on CIFAR10. Bold numbers highlight the best results.

Networks	Models	Type	Clean	FGSM	BIM	PGD 10	PGD 20	PGD 50	C&W	AA
ResNet34	TDAT	best	82.41	66.38	56.38	57.65	56.94	56.73	49.18	46.17
		final	83.08	66.23	55.63	56.64	55.80	55.48	49.20	45.99
	Ours	best	**83.16**	**68.76**	**60.57**	**61.47**	**60.96**	**60.91**	**49.37**	**47.15**
		final	**83.16**	**68.76**	**60.57**	**61.47**	**60.96**	**60.91**	**49.37**	**47.15**
VGG11	TDAT	best	76.49	59.58	50.51	50.88	50.07	49.93	**43.26**	**41.20**
		final	76.63	59.66	50.34	50.85	50.13	49.88	**43.13**	**41.16**
	Ours	best	**76.66**	**62.86**	**54.44**	**54.99**	**54.46**	**54.19**	42.49	40.44
		final	**76.66**	**62.86**	**54.44**	**54.99**	**54.46**	**54.19**	42.49	40.44

Table 4. Robustness under varied budgets.. Bold numbers highlight the best results.

Methods	Eval Budgets	Clean	FGSM	BIM	PGD 10	PGD 20	PGD 50	C&W	AA
FGSM-PGK	$\frac{12}{255}$	81.78	56.94	41.30	43.51	39.38	38.30	33.07	**28.86**
TDAT		**82.60**	**60.41**	**44.33**	**47.44**	**43.55**	**42.60**	**33.08**	27.91
FGSM-PGK	$\frac{11}{255}$	81.78	59.11	44.52	45.95	43.27	42.75	37.02	**33.36**
TDAT		**82.60**	**62.66**	**48.12**	**49.94**	**47.28**	**46.62**	**37.02**	32.23
FGSM-PGK	$\frac{10}{255}$	81.78	61.26	48.11	48.97	47.34	47.00	41.16	**37.89**
TDAT		**82.60**	**64.92**	**51.86**	**53.15**	**51.43**	**50.95**	**41.47**	36.81
FGSM-PGK	$\frac{9}{255}$	81.78	63.34	52.20	52.87	51.61	51.23	45.66	**43.31**
TDAT		**82.60**	**66.81**	**55.85**	**56.88**	**55.80**	**55.46**	**45.71**	42.52
FGSM-PGK	$\frac{7}{255}$	81.78	67.57	59.97	60.24	59.93	59.76	**54.63**	**53.01**
TDAT		**82.60**	**70.55**	**63.03**	**63.56**	**63.15**	**62.96**	54.23	51.49
FGSM-PGK	$\frac{6}{255}$	81.78	69.86	63.61	63.75	63.57	63.53	**59.15**	**57.56**
TDAT		**82.60**	**72.42**	**66.47**	**66.83**	**66.66**	**66.58**	58.85	56.43
FGSM-PGK	$\frac{5}{255}$	81.78	72.27	67.15	67.21	67.09	67.04	63.28	**62.19**
TDAT		**82.60**	**74.20**	**69.66**	**69.88**	**69.82**	**69.72**	**63.34**	61.54
FGSM-PGK	$\frac{4}{255}$	81.78	74.46	70.77	70.85	70.79	70.79	67.17	**66.32**
TDAT		**82.60**	**76.03**	**72.51**	**72.71**	**72.65**	**72.51**	**67.59**	66.11

class plays a crucial role in maintaining model stability.

Ablation studies on α in Eq. (9). α controls the additional regularization strength applied to misclassified samples. We vary α from 0 to 2.1 with a step size of 0.3, and report the results in Fig. 2. As α increases, the clean accuracy gradually decreases, while the robustness under various adversarial attacks consistently improves. However, excessively large α values lead to unstable training. Based on this, we set $\alpha = 1.5$ as the default configuration in our work.

Ablation studies on κ in Eq. (5). To investigate the impact of perturbation magnitude, we vary the scaling factor κ from 1 to 4 and analyze its effect on model performance. κ serves as a scaling factor controlling the amplitude of dynamic variation. As illustrated in Figure 3, increasing κ leads to a steady improvement in clean accuracy, while adversarial robustness reaches its optimum at $\kappa = 2$ and then gradually declines. This degradation is primarily attributed to the insufficient perturbation applied to low-confidence samples, which limits their adversarial exposure and weakens robustness learning.

Illustration of confidence grouping. For the catastrophic overfitting analysis, we divide each training batch into four confidence groups. The primary motivation is that this analysis focuses on model stability, and assigning more samples to larger adversarial perturbations increases the risk of catastrophic overfitting. For the performance-tradeoff analysis, we split each batch into 32 groups. On one hand, analyzing performance trade-offs requires ensuring stable training; on the other hand, a finer partition provides more detailed insights into the trade-off behavior. For analyzing the effect of the supervision signal, we divide samples into 8 groups, based on the model's prediction state. On CIFAR-10, the clean accuracy of the model does not exceed 85%. Thus, most samples in the lowest confidence group correspond to incorrectly classified instances.

Illustration of many hyperparameters. To clearly analyze the effects of different factors, we introduce several hyperparameters, some of which have clear tuning patterns

or limited impact on performance. 1) We adopt the default value of γ in TDAT. 2) As shown in Fig. 4 of the main paper, λ exhibits a slight impact on model performance. 3) As shown in Figs. 2 and 2 of this supplement, α and κ exhibit clear patterns: increasing κ or decreasing α improves clean accuracy but degrades adversarial robustness; Moreover, the perturbation magnitude controlled by κ has explicit bounds. 4) τ_1 aims to select high-confidence misclassified samples, and the ratio τ_1/B is smaller than the misclassification rate. Therefore, the hyperparameter tuning is typically performed only over α and τ_1.

Marginal gains on AutoAttack. 1) AutoAttack (especially APGD-T) reveals a common bottleneck against exhaustive targeted attacks in FAT—not a specific shortcoming of DDG. APGD-T targets all non-ground-truth classes (NCs), exploiting the "weakest link" in the model's decision boundary. However, suppressing the confidence of all NCs is highly challenging. Hence, multiple methods (FGSM-PGK, FGSM-PGI, TDAT, and our DDG) exhibit similar AutoAttack results. 2) This highlights our work's positioning: DDG achieves superior robustness across a broader spectrum of attacks (FGSM, PGD, APGD) through a fundamental cognitive bias correction mechanism.

Difference from Prior Work. Huang et al. [1] introduced a fast adversarial training method with adaptive step sizes, where each sample's step size is inversely proportional to the regularized gradient of the loss with respect to the input, aiming to maintain similar adversarial strengths across samples. In contrast, we argue that enforcing uniform adversarial enhancement across all samples is suboptimal. We propose a dynamic perturbation budget allocation strategy that adaptively adjusts perturbation strength, encouraging samples to converge toward consistent decision boundaries. Additionally, Liu et al. [2] developed an adaptive-guided adversarial training framework that combines MSE and RMSE losses with a stop-gradient operation. Different from their focus on loss design, our method jointly introduces dynamic perturbation budgets and adaptive supervision signals, effectively preventing excessive emphasis on misclassified samples.

Illustration of Eq. (6). Eq .(6) is formulated as:

$$\delta_i = \mathrm{clip}\left(\delta_{\text{init}} + \max\{\xi_i, \xi_{\text{base}}\} \cdot \mathrm{sign}(\nabla_{x_i+\delta_{\text{init}}}\mathcal{L}),\ -\xi_i,\ \xi_i\right)$$

If the $\max\{\xi_i, \xi_{\text{base}}\}$ operation is omitted, the formulation simplifies to:

$$\delta_i = \mathrm{clip}\left(\delta_{\text{init}} + \xi_i \cdot \mathrm{sign}(\nabla_{x_i+\delta_{\text{init}}}\mathcal{L}),\ -\xi_i,\ \xi_i\right)$$

Here, δ_{init} typically takes fixed perturbation values, such as $-8/255$ or $8/255$. The effect of gradient direction on the perturbation update is analyzed below under two scenarios.

Case 1: $|\xi_i| \leq 8/255$

- Without the $\max$ operator: (1) When the gradient direction of ξ_i aligns with that of δ_{init}, the clipping retains a perturbation of magnitude ξ_i in the same direction as δ_{init}. (2) When the gradient directions oppose each other, the clipping preserves a perturbation of magnitude $8/255 - \xi_i$, still aligned with δ_{init}'s gradient. In this case, the update always inherits the gradient direction of δ_{init}, which may not be efficient for adversarial enhancement.

- With $\max\{\xi_i, \xi_{\text{base}}\}$: The gradient direction of δ_{init} is retained only when it aligns with that of ξ_i.

Case 2: $|\xi_i| > 8/255$

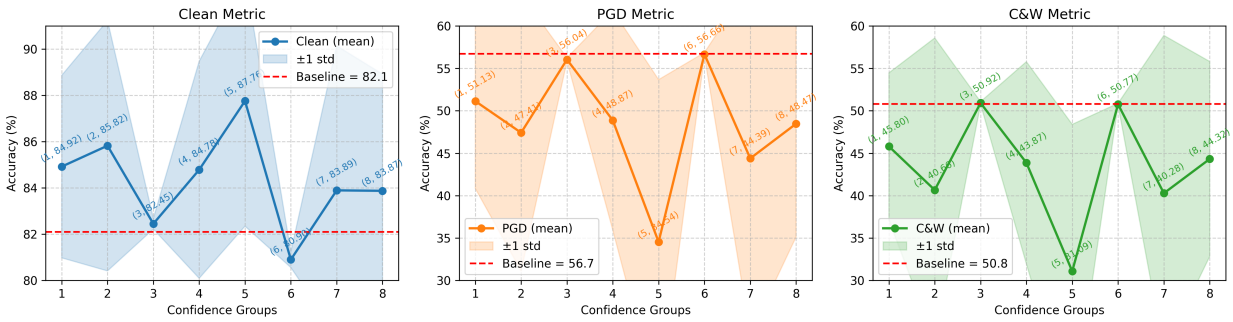
With or without the $\max$ operator: (1) If the gradient directions of ξ_i and δ_{init} are aligned, the clipping retains a perturbation of magnitude ξ_i in the direction of δ_{init}. (2) If the gradient directions are opposed, the clipping retains a perturbation of magnitude $\xi_i - 8/255$ in the direction of ξ_i's gradient. - Thus, the update always follows the gradient direction of ξ_i, effectively suppressing the influence of δ_{init}.

The introduction of the $\max\{\xi_i, \xi_{\text{base}}\}$ operator prevents reliance on potentially inefficient initial gradient directions when $|\xi_i|$ is small. This ensures that the perturbation updates consistently follow an adversarial direction, thereby improving training stability and adversarial robustness.

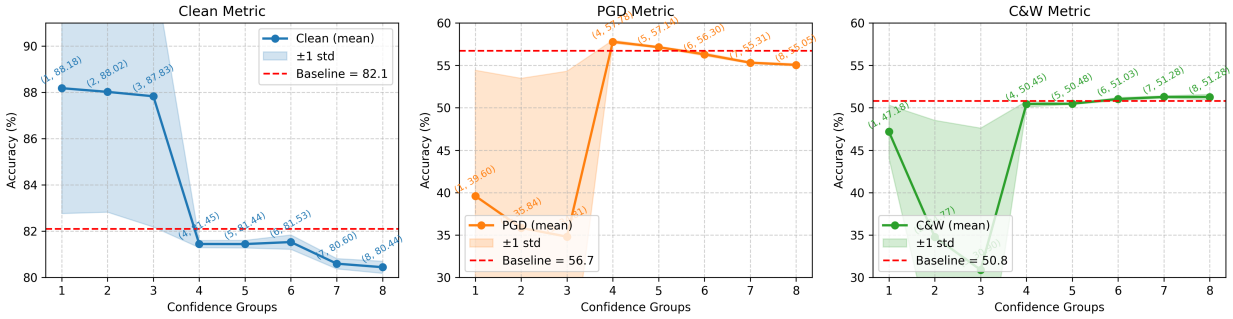
References

[1] Zhichao Huang, Yanbo Fan, Chen Liu, Weizhong Zhang, Yong Zhang, Mathieu Salzmann, Sabine Süsstrunk, and Jue Wang. Fast adversarial training with adaptive step size. *TIP*, 32:6102–6114, 2023. 3

[2] Zhenyu Liu, Huizhi Liang, Xinrun Li, Vaclav Snasel, and Varun Ojha. Adagat: Adaptive guidance adversarial training for the robustness of deep neural networks. *arXiv preprint arXiv:2508.17265*, 2025. 3



(a) Setting β_1 to -0.1 and β_2 to 0.



(b) Setting β_1 to 0 and β_2 to 0.1.

Figure 1. Impact of supervision strength on CIFAR10 and ResNet18.

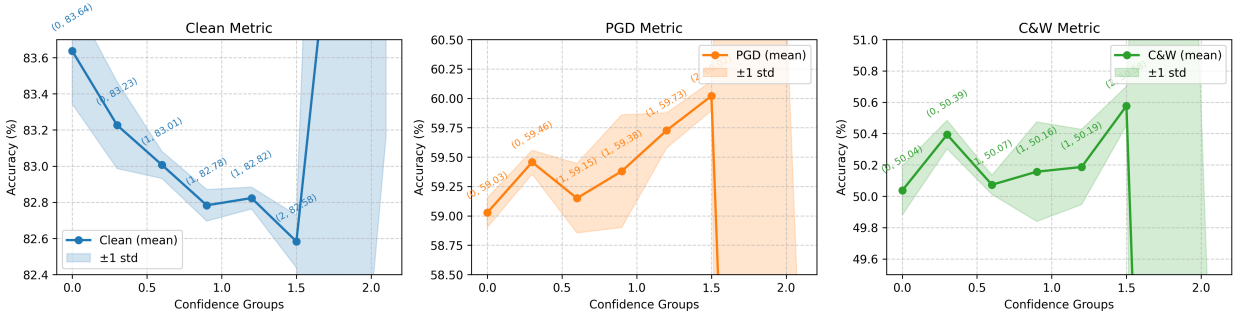


Figure 2. Impact of the hyperparameter α on CIFAR10 and ResNet18. Both PGD and C&W are with 10 steps.

Figure 3. Impact of the hyperparameter κ on CIFAR10 and ResNet18. Both PGD and C&W are with 10 steps.