Supplementary Material for "Fast Adversarial Training with Smooth Convergence"

Mengnan Zhao, Lihe Zhang*, Yuqiu Kong and Baocai Yin

Dalian University of Technology

1. Detailed algorithms of the proposed methods

The details of our proposed example-based ConvergeSmooth, batch-based ConvergeSmooth, and weight centralization are shown in Algorithms 1, 2, and 3 respectively. Other attack initialization methods can be easily combined with ConvergeSmooth by replacing line 15 in Algorithm 1.

2. Additional Experiments

Graphical analysis of various models and datasets. Fig. 1 of the main text shows the training process of adversarially trained models (ResNet18) using previous FAT methods on the CIFAR10 dataset. We then provide the graphical analysis of the CIFAR10 with WideResnet, as well as the CIFAR100 with Resnet18 in this supplementary material. Results are given in Fig. 1. These experimental results also verify our conclusions, *e.g*. the mentioned FAT methods face catastrophic overfitting under the perturbation budget 16/255. See Section 3.2 of the main text for a detailed analysis.

We discover that the default FGSM-RS overfits within 10 epochs at ξ = 16/255. Removing constraint $\delta_0 + \delta \in [-\xi, \xi]$ improves the diversity of perturbations and brings a significant improvement. The training processes with or without using this constraint are shown in Fig. 2.

Various Networks. The main text studies the adversarial robustness of FAT methods on ResNet18 [6]. Here, we adopt WideResNet34 with a width factor of 10 [12] as the backbone. WideResNet34-10 is more complex than ResNet18 and takes much more time to train. The experimental results on the CIFAR-10 and CIFAR-100 datasets are given in Tabs. 1 and 2, respectively. We note that our proposed ConvergeSmooth prevents the wider architectures from the catastrophic overfitting problem. It shows higher adversarial robustness than all other FAT methods and comparable performance to PGD-AT with less time consumption.

Experiments on ImageNet We also conduct experiments on ImageNet [3]. The initial learning rate is set to 0.1, ResNet50 [6] is selected as the backbone, and FGSM-BP [11] is adopted as the initialization method. Then, we optimize models with a total training epoch of 90 and decay the learning rate at the 30_{th} and 60_{th} epoch with a factor of 0.1. For hyper-parameters, γ_{max} = 0.045, w_1 = 0 and w_2 = 1. The experimental results are given in Tab. 3. We observe that a stable FAT training with ξ = 16/255 on ImageNet struggles to converge. Overall, our proposed method can stabilize the FAT process on various datasets.

Analysis of the learning rate and early stopping. [8] trains the model with the cyclic learning rate (CLR) and early stopping (ES). However, the setting of CLR is not always available for training FAT methods under large perturbation budgets. For instance, the retrained FGSM-RS model achieves 83.2 and 43.7 on clean and PGD50 metrics at 8/255 with the CLR and 15 epochs. Under the same settings, the adversarial robustness of FGSM-RS drops to 0 at 16/255.

Moreover, ES requires evaluating the model after each training epoch to determine whether early stopping should be utilized, which causes excessive time consumption. Premature early stopping seriously damaged the performance of models. For example, the default FGSM-RS faces catastrophic overfitting at the 8th training epoch under the setting of the multistep learning rate (30 epochs decay at 20 and 25) and ξ = 16/255. The final evaluation result is 60.1 and 15.2 on clean and AA accuracy. Under the same settings, our B-RS realizes 66.8 and 18.8 on clean and AA accuracy.

*Corresponding author: zhanglihe@dlut.edu.cn

Algorithm 1: Example-based ConvergeSmooth-RS.

Input: The epoch N, the dataset D, the cross-entropy loss $\mathcal{L}$, the model $f(\cdot;\theta)$, the hyper-parameters w_1, w_2, γ_{max} and γ_{min}, the perturbation budget ξ.

Output: The adversarially trained model $f(\cdot;\theta)$.

for t *in* N **do**
  $\mathcal{L}_{adv}^{sum}$ = 0;
  $\mathcal{L}_{ori}^{sum}$ = 0;
  Iter = 0;
  for (x_0, y) *in* D **do**
    $\mathcal{L}_{ori} = \mathcal{L}(f(x_0;\theta), y)$;
    $\mathcal{L}_{ori}^{sum}$ += $\mathcal{L}_{ori}$;
    Iter += 1;
    $\mathcal{L}_{CS}$ = 0;
    if $t > 2$ **then**
      if $|\mathcal{L}_{ori} - u_{t-1}| > \gamma_t$ **then**
        $\mathcal{L}_{CS} = w_1 \cdot \mathcal{L}_{adv} \cdot \text{sgn}(\mathcal{L}_{adv} - u'_{t-1}) + w_2 \cdot \mathcal{L}_{ori} \cdot \text{sgn}(\mathcal{L}_{ori} - u_{t-1})$;
      end
    end
    $\delta_0 = \mathbf{U}(-\xi, \xi)$;
    $g_c = \text{sgn}(\nabla_{x_0+\delta_0} \mathcal{L}(f(x_0 + \delta_0;\theta), y))$;
    $\delta = \text{clip}_{\xi}[\delta_0 + \xi \cdot g_c]$;
    $\delta = \text{clip}_{0.5}[x_0 + \delta - 0.5]$;
    $\mathcal{L}_{adv} = \mathcal{L}(f(x_0 + \delta;\theta), y)$;
    $\mathcal{L}_{ori}^{sum}$ += $\mathcal{L}_{adv}$;
    $\mathcal{L}_{adv}$ += $\mathcal{L}_{CS}$
    $\theta = \theta - \nabla_\theta \mathcal{L}_{adv}$;
  end
  $u_{t-1} = \mathcal{L}_{ori}^{sum}$ / Iter;
  $u'_{t-1} = \mathcal{L}_{adv}^{sum}$ / Iter;
  if $t > 2$ **then**
    $d_{t-1} = u_{t-1} - u_{t-2}$;
    $\gamma_t = \min(\max(d_{t-1}, \gamma_{min}), \gamma_{max})$;
  end
end

For fair comparisons, we follow FGSM-MEP to train all methods with a multistep learning rate (MLR) and 110 epochs (decay at 100 and 105) without considering early stopping (ES). Hyperparameters are tuned for training previous methods on a large perturbation budget.

Different hyperparameter settings. Parameter selection suggestion: set initial γ_{max} = 0.03, increase γ_{max} to improve the attack performance if the FAT process is stable, otherwise increase the value of w_1 to make the FAT process stable. Other hyperparameters choose default values.

There are 4 hyperparameters in ConvergeSmooth, γ_{min}, γ_{max}, w_1 and w_2. w_1 and w_2 are usually set to 0 and 1 respectively. w_1 is gradually increased from 0.3 with a stride of 0.2 once the training process faces catastrophic overfitting (w_1 = 0 and w_2 = 1 by default). By sacrificing some robustness, γ_{max} and γ_{min} can be replaced by γ. Hence, we mainly adjust w_1 and γ. γ is used to prevent a small amount of data from overfitting and is closely related to the loss difference of adjacent epochs. Models exhibit different loss variances on various datasets. In Tab. 4, we observe that too small γ will affect performance and too large γ cannot prevent overfitting. Meanwhile, the classification accuracy of benign samples and adversarial samples in stable training is not sensitive to the hyperparameter settings. The specific details of hyperparameter settings in this work are shown in Tab. 5.

Analysis of loss types. In the main text of this work, the L_1 loss and dynamic convergence stride are utilized to control the history difference and convergence speed, respectively. Next, we replace the L_1 loss and dynamic convergence stride

Algorithm 2: Batch-based ConvergeSmooth-RS.

Input: The epoch N, the dataset D, the cross-entropy loss $\mathcal{L}$, the model $f(\cdot;\theta)$, the hyper-parameters w_1, w_2, γ_{max} and γ_{min}, the perturbation budget ξ.

Output: The adversarially trained model $f(\cdot;\theta)$.

for t *in* N **do**
  $\mathcal{L}_{adv}^{sum} = 0$;
  $\mathcal{L}_{ori}^{sum} = 0$;
  Iter = 0;
  /* For a batch of data */
  for *(B, T) in D* **do**
    $\mathcal{L}_{ori} = \mathcal{L}(f(B;\theta), T)$; /* The average loss of batch data */
    $\mathcal{L}_{ori}^{sum}$ += $\mathcal{L}_{ori}$;
    Iter += 1;
    $\mathcal{L}_{CS} = 0$;
    if $t > 2$ **then**
      if $|\mathcal{L}_{ori} - u_{t-1}| > \gamma_t$ **then**
        $\mathcal{L}_{CS} = w_1 \cdot \mathcal{L}_{adv} \cdot \text{sgn}(\mathcal{L}_{adv} - u'_{t-1}) + w_2 \cdot \mathcal{L}_{ori} \cdot \text{sgn}(\mathcal{L}_{ori} - u_{t-1})$;
      end
    end
    $\delta_0 = \mathbf{U}(-\xi, \xi)$; /* The initialization perturbations of batch data */
    $g_c = \text{sgn}(\nabla_{B+\delta_0} \mathcal{L}(f(B+\delta_0;\theta), T))$;
    $\delta = \text{clip}_{\xi}\ [\delta_0 + \xi \cdot g_c]$;
    $\delta = \text{clip}_{0.5}\ [\delta + B - 0.5]$;
    $\mathcal{L}_{adv} = \mathcal{L}(f(B+\delta;\theta), T)$; /* The average loss of batch data */
    $\mathcal{L}_{ori}^{sum}$ += $\mathcal{L}_{adv}$;
    $\mathcal{L}_{adv}$ += $\mathcal{L}_{CS}$
    $\theta = \theta - \nabla_\theta \mathcal{L}_{adv}$;
  end
  $u_{t-1} = \mathcal{L}_{ori}^{sum}$ / Iter;
  $u'_{t-1} = \mathcal{L}_{adv}^{sum}$ / Iter;
  if $t > 2$ **then**
    $d_{t-1} = u_{t-1} - u_{t-2}$;
    $\gamma_t = \min(\max(d_{t-1}, \gamma_{min}), \gamma_{max})$;
  end
end

with the L_2 loss and exponential moving average (EMA) respectively. Meanwhile, EMA is also applied to update the model weights θ_t. Tab. 6 provides detailed results. It can be seen that the model is not sensitive to the choice of loss functions.

Compare with the AT method. 1) Our methods are proposed to solve the catastrophic overfitting problem in FAT process. It has been proved in the main text that AT methods are computationally expensive but stable, so it is not necessary to apply additional constraints to AT methods since there is no catastrophic overfitting problem. Thus, we apply the AT method such as TRADES [13] to FAT. 2) TRADES and NuAT share a similar implementation in FAT, *i.e.* TRADES and NuAT only have different initialization perturbations and loss functions for gradient backpropagation. Mainstream initialization perturbations follow uniform distribution (UD) or random distribution (RD). $1/\lambda$ in TRADES is set to 10 or 5. In Tab. 7, the results on CIFAR-100 and the perturbation budget 12/255 are given. ResNet-18 is adopted as the backbone. Experiments show that the AT method may not be effective in the FAT task.

Various perturbation budgets. In addition to the experiments for the perturbation budgets 10/255, 12/255, and 16/255, similar experiments are performed for ξ= 8/255. The results in Tab. 8 prove that our approach does not reduce classification performance (in fact it improves slightly).

The difference. SAF (or MESA) [4] separately flattens the logit of the corresponding sample pair between epochs. Besides inter-epoch flattening, our ConvergeSmooth further synchronizes loss predictions of all samples within the same

Algorithm 3: Weight Centralization.

Input: The epoch N, the training dataset D, the cross-entropy loss $\mathcal{L}$, the model $f(\cdot;\theta_0)$, the hyper-parameters w_3, the perturbation budget ξ, the flag $eval$ ('False' by default).

Output: The adversarially trained model $f(\cdot;\theta_N)$.

for t *in* N **do**
  Iter = 0;
  SumAcc = 0;
  /* For a batch of data */
  for *(B, T) in D* **do**
    $\mathcal{L}_{CS}$ = 0;
    if $t > 2$ **then**
      $\mathcal{L}_{CS} = w_3 \cdot ||\theta_t - \frac{\theta^*}{Iter}||_p$;
    end
    δ_0 = $\mathbf{U}(-\xi, \xi)$; /* The initialization perturbations of batch data */
    g_c = sgn($\nabla_{B+\delta_0} \mathcal{L}(f(B+\delta_0;\theta), T)$);
    δ = clip$_\xi$ [δ_0 + $\xi \cdot g_c$];
    δ = clip$_{0.5}$ [δ + B - 0.5];
    $\mathcal{L}_{adv}$ = $\mathcal{L}(f(B+\delta;\theta), T)$; /* The average loss of batch data */
    $\mathcal{L}_{adv}$ += $\mathcal{L}_{CS}$;
    θ_t = θ_t - $\nabla_{\theta_t}\mathcal{L}_{adv}$;
  end
  if $eval$ **then**
    /* D_1 is a randomly selected subset from the validation dataset */
    if $PGD\_Acc(D_1, \theta_t) \geq \frac{Sum\_Acc}{Iter}$ **then**
      θ^* += θ_t;
      Sum_Acc += $PGD\_Acc(D_1, \theta_t)$;
      Iter += 1;
    end
  else
    θ^* += θ_t;
    Iter += 1;
  end
end

epoch, solving catastrophic overfitting. 2) SAF causes the memory overload issue in large datasets. Ours does not. 3) MEA used in MESA is verified to be ineffective in the FAT task, please see Table 6 of the supplement. (4) [2] enforces the model loss increase with the perturbation size. Our models are trained under a fixed perturbation size.

Additional experiments. Tabs. 9-11 provide a more comprehensive assessment of the results under the settings of the main text, covering the average evaluation results of the best model across the three runs (*mbest*), the best results (*best*) and the average evaluation of the final model from the three runs (*mfinal*).

References

[1] Flammarion N Andriushchenko M. Understanding and improving fast adversarial training. In *Advances in Neural Information Processing Systems*, pages 16048–16059, 2020.

[2] RV Babu BS Vivek. Regularizers for single-step adversarial training. In *arXiv preprint arXiv:2002.00614*, 2020.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. *Advances in Neural Information Processing Systems*, 35:23439–23451, 2022.

[5] Saberi M. Eskandar M. Golgooni, Z. and M. H Rohban. Zerograd: Mitigating and explaining catastrophic overfitting in fgsm adversarial training. page arXiv preprint arXiv:2103.15476, 2021.

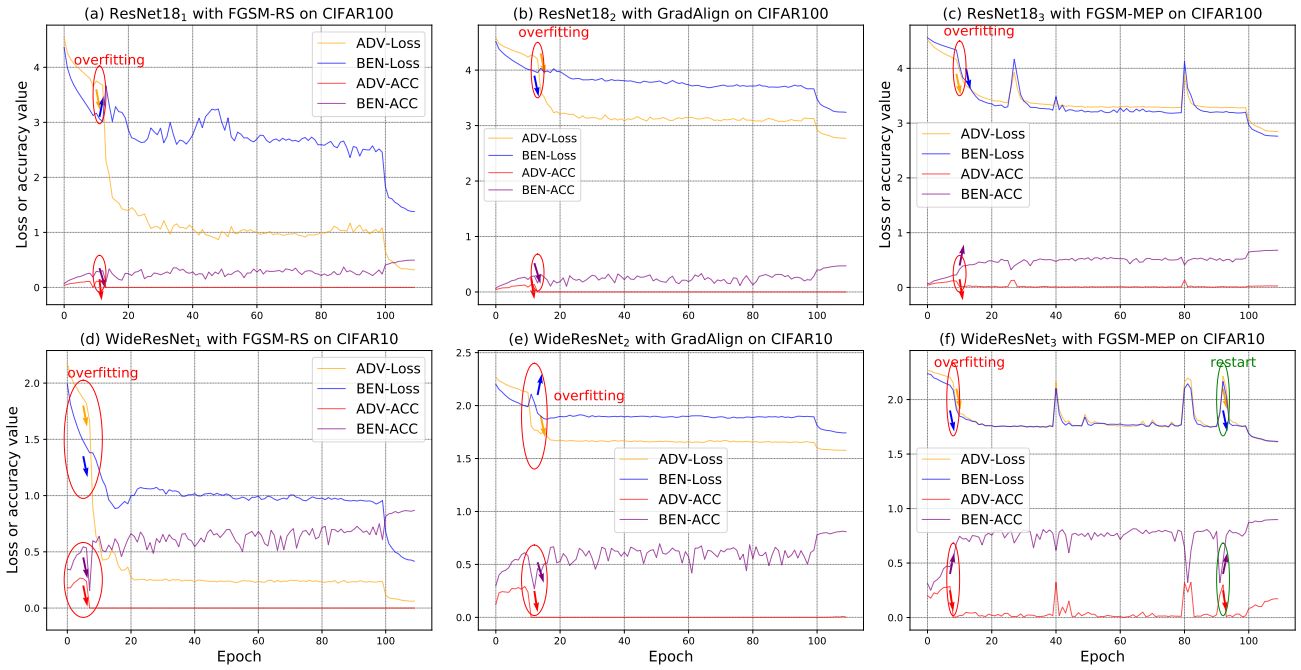

Figure 1. Graphical analysis of various models and datasets.

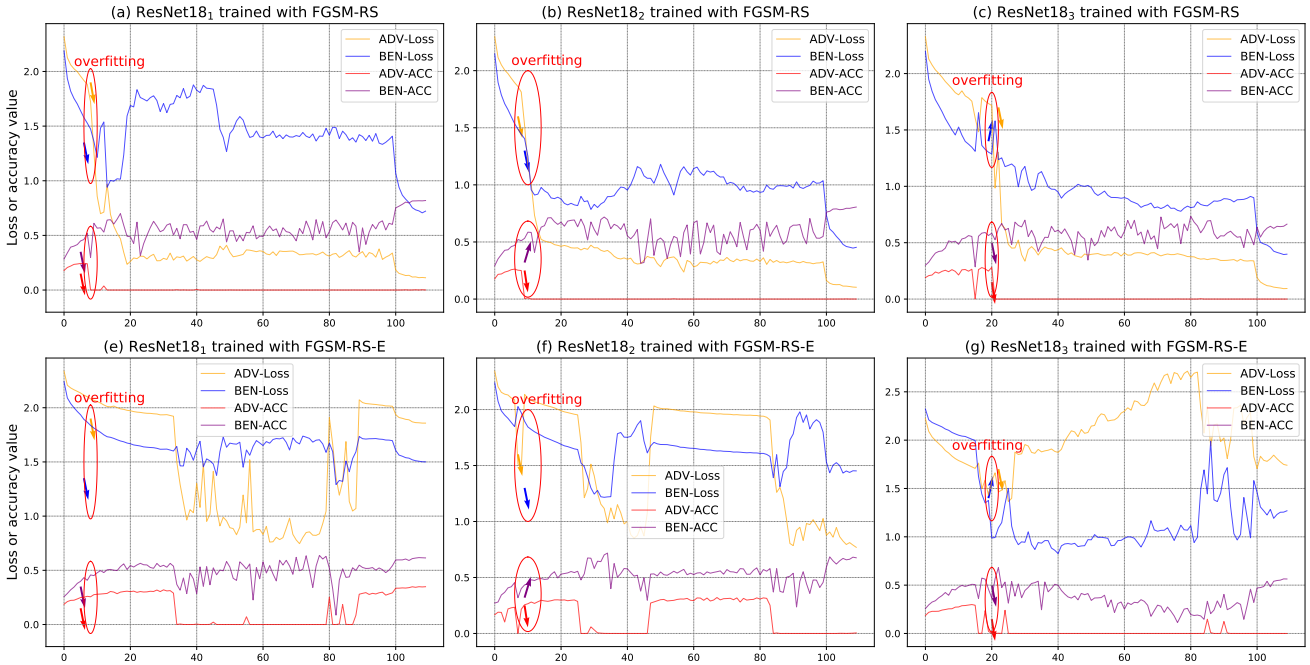

Figure 2. Ablation studies for the constraint $\delta_0 + \delta \in [-\xi, \xi]$ in FGSM-RS. -E means removing the constraint $\delta_0 + \delta \in [-\xi, \xi]$.

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

[7] Zhichao Huang, Yanbo Fan, Chen Liu, Weizhong Zhang, Yong Zhang, Mathieu Salzmann, Sabine Süsstrunk, and Jue Wang. Fast adversarial training with adaptive step size. *arXiv preprint arXiv:2206.02417*, 2022.

[8] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

[9] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, et al. Towards efficient and effective adversarial training. volume 34, pages

Methods		Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD↑	AA ↑	Time (hours)↓
PGD-AT [8]	*best*	70.76	52.82	44.08	37.45	35.59	31.13	34.02	26.41	26.1
	final	70.76	52.82	44.08	37.45	35.59	31.13	34.02	26.41	
FGSM-RS [10]	*best*	53.94	32.32	27.60	22.24	20.96	21.17	20.27	16.28	4.49
	final	78.96	74.25	2.64	0.71	0.16	0.06	0.37	0.00	
GradAlign [1]	*best*	55.58	34.10	27.22	21.49	19.92	17.56	18.97	13.45	8.79
	final	71.67	78.16	1.23	0.17	0.01	0.12	0.00	0.00	
ZeroGrad [5]	*best*	74.25	46.74	35.62	25.05	21.15	23.61	18.55	13.79	4.49
	final	85.72	59.41	23.92	13.13	7.96	13.18	3.98	3.08	
NuAT [9]	*best*	74.25	46.74	35.62	25.05	21.15	23.61	18.55	13.19	7.66
	final	84.82	69.98	5.98	1.96	0.69	1.67	0.12	0.01	
ATAS[7]	*best*	79.10	47.66	34.70	23.89	20.36	23.14	18.67	14.86	5.25
	final	83.32	58.26	26.07	16.28	13.18	17.17	10.19	8.58	
FGSM-MEP [11]	*best*	65.56	38.68	32.51	26.15	25.0	18.2	24.29	15.43	6.15
	final	**90.56**	**83.25**	11.52	5.95	2.57	1.98	1.36	0.01	
Ours-B-MEP	*best*	69.94	51.17	**43.34**	**36.75**	**34.79**	**29.69**	**32.83**	**24.27**	6.72
	final	71.22	51.32	42.70	35.70	33.65	29.33	31.52	23.51	

Table 1. Quantitative results of the adversarial training methods (ξ = 16/255) on CIFAR-10 with WideResNet as the backbone.

Methods		Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD↑	AA ↑	Time (hours)↓
PGD-AT [8]	*best*	46.06	25.98	22.76	18.32	17.60	15.24	17.29	12.97	26.1
	final	49.83	27.01	21.81	17.22	16.55	14.76	16.27	12.44	
FGSM-RS [10]	*best*	25.89	15.59	13.34	11.26	10.98	9.18	10.8	7.82	4.49
	final	41.17	33.85	0.00	0.00	0.00	0.00	0.00	0.00	
GradAlign [1]	*best*	35.93	19.62	15.30	11.57	10.48	8.96	10.21	7.13	8.79
	final	48.60	46.24	0.00	0.00	0.00	0.00	0.00	0.00	
ZeroGrad [5]	*best*	51.31	26.18	18.46	12.83	11.09	12.41	10.06	7.59	4.49
	final	63.15	41.98	1.46	0.55	0.15	0.24	0.00	0.00	
NuAT [9]	*best*	20.48	13.88	12.37	10.93	10.73	8.56	10.31	7.30	7.66
	final	69.99	33.54	5.90	3.34	2.11	3.17	1.14	0.69	
ATAS[7]	*best*	71.55	40.26	**24.31**	16.44	9.80	10.91	1.79	0.03	5.25
	final	71.95	41.21	21.02	14.03	7.71	8.52	0.72	0.00	
FGSM-MEP [11]	*best*	20.69	12.64	11.14	9.68	9.52	7.67	9.46	6.73	6.15
	final	**72.12**	**58.57**	5.30	2.68	1.22	0.54	1.03	0.00	
Ours-B-MEP	*best*	48.64	26.74	22.27	17.70	**16.95**	14.65	**16.50**	**12.05**	6.72
	final	48.45	26.72	22.26	**17.79**	16.90	**14.71**	16.43	11.86	

Table 2. Quantitative results of the adversarial training methods (ξ = 16/255) on CIFAR-100 with WideResNet as the backbone.

Methods	Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD↑	AA ↑	Time (hours)↓
FGSM-RS [10]	4.83	2.48	2.48	1.67	1.24	0.82	0.56	0.19	56.2
FGSM-BP[11]	13.04	8.68	1.70	0.59	0.26	0.47	0.22	0.08	75.7
Ours-B-BP	**26.18**	**14.53**	**8.10**	**4.01**	**1.96**	**2.57**	**1.49**	**0.59**	82.5

Table 3. Quantitative results of the adversarial training methods (ξ = 16/255) on ImageNet with ResNet50 as the backbone. The digit denotes the optimal adversarial robustness of the model against the PGD10 attack.

11821–11833, 2021.

[10] Kolter J Z. Wong E, Rice L. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020.

[11] Xingxing Wei Baoyuan Wu Ke Ma Jue Wang Xiaochun Cao Xiaojun Jia, Yong Zhang. Prior-guided adversarial initialization for fast adversarial training. In *Proceedings of the European conference on computer vision (ECCV)*, 2022.

[12] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Dataset	γ	Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD↑	AA ↑	Stability
CIFAR10	0.03	72.63	54.40	45.23	42.85	42.14	36.81	41.62	33.26	⋆⋆⋆
	0.045	72.96	54.88	45.54	42.89	42.19	37.52	40.50	32.86	⋆⋆
	0.06	87.92	82.21	20.74	15.37	11.62	9.36	7.25	0.39	-
CIFAR100	0.03	48.13	32.13	24.25	22.67	22.21	19.04	21.29	15.28	⋆⋆⋆
	0.045	48.19	31.56	24.49	22.68	22.29	19.01	21.26	15.54	⋆⋆⋆
	0.06	49.86	32.43	24.79	23.12	22.78	19.52	21.10	15.51	⋆⋆

Table 4. Quantitative results of the proposed method on various γ with ResNet18 as the backbone and the perturbation budget 12/255. CIFAR10 and CIFAR100 are selected as datasets. 'Stability' represents the number of times the model is stable in three training repetitions. w_1 and w_2 are set to 0 and 1, respectively.

FAT methods	Datasets	γ_{max}	w_1	w_2	w_3
Ours-E-MEP	CIFAR10,100	0.06	0	1	-
Ours-E-RS	CIFAR10	0.06	0	1.5	-
	CIFAR100	0.06	0	0.5	-
Ours-B-MEP	CIFAR10	0.03	0	1	-
	CIFAR100	0.06	0	1	-
	Tiny-ImageNet	0.03	0.5	1	-
	ImageNet	0.045	0	1	-
Ours-B-RS	CIFAR10	0.03	0	1	-
	CIFAR100	0.06	0	1	-
Ours-W-RS	CIFAR10,100	-	-	-	0.1

Table 5. Specific details of hyperparameter settings in this paper.

Methods	Ours-B-MEP	L_1+EMA			L_2 (w_2=0.1)	WEMA		
Decay	-	0.1	0.5	0.9	-	0.1	0.5	0.9
w_2	1.0	1.0	1.0	1.0	0.1	1.0	1.0	1.0
Clean/AA	49.9/15.5	32.9/10.8	44.8/15.2	47.0/13.6	48.3/15.3	30.9/10.4	32.9/10.7	37.7/14.3

Table 6. Ablation studies for various loss types and strategies with ResNet18 as the backbone and the perturbation budget 12/255. CIFAR100 is selected as the dataset. 'Ours' means training with L_1 loss and w_2 = 1. 'WEMA' denotes that we use EMA to update the current model weights.'Decay' is set to 0.1, 0.5, or 0.9.

Methods	Distribution	$1/\lambda$	Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD↑	AA ↑
TRADES	UD	5.0	63.51	24.44	12.72	9.92	9.17	8.72	8.67	7.30
		10.0	62.88	26.54	15.62	12.73	11.97	10.72	11.03	9.01
	RD	5.0	48.24	18.64	10.26	8.58	8.12	7.64	7.63	6.15
		10.0	44.49	20.68	12.60	11.12	10.73	9.30	10.23	7.71
Ours-B-MEP	RD	-	49.86	32.43	24.79	23.12	22.78	19.52	21.10	15.51

Table 7. Comparitive experiments between the AT method and the proposed method on ResNet18 as the backbone and the perturbation budget 12/255. CIFAR100 is selected as the dataset.

[13] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

ξ	FAT methods		Clean↑	PGD-10↑	PGD-50↑	AA
8/255	FGSM-RS	*best*	73.81	42.31	41.26	37.69
		final	**83.82**	0.09	0.02	-
	Ours-B-RS	*best*	82.18	51.94	50.50	44.13
		final	82.25	51.80	50.42	-
	FGSM-MEP	*best*	81.72	55.22	54.19	49.00
		final	82.05	55.10	54.08	-
	Ours-B-MEP	*best*	81.56	**55.92**	**54.98**	49.34
		final	81.67	55.53	54.54	-

Table 8. Quantitative results of FAT methods on classical ξ with ResNet18 as the backbone and CIFAR-10 as the dataset. Models are trained and evaluated under the same ξ. The number in bold indicates the best.

Methods		Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD↑	AA ↑	Time (min)↓
PGD-AT [8]	*best*	65.32	46.43	40.83	35.30	34.09	30.86	33.23	26.39	370
	mbest	65.30	46.31	40.73	35.08	33.92	30.84	33.08	26.29	
	mfinal	65.29	46.28	40.79	35.06	33.77	30.31	33.01	25.92	
FGSM-RS [10]	*best*	46.99	36.33	25.96	21.88	21.16	19.73	20.73	16.59	67
	mbest	50.12	38.28	26.13	21.55	20.43	18.96	19.40	14.84	
	mfinal	76.31	50.67	0.00	0.00	0.00	0.00	0.00	0.00	
GradAlign [1]	*best*	59.03	39.31	33.54	27.68	26.14	21.46	25.18	17.67	135
	mbest	58.17	39.87	33.12	26.81	24.99	22.63	23.98	17.02	
	mfinal	70.86	69.51	0.00	0.00	0.00	0.00	0.00	0.00	
ZeroGrad [5]	*best*	74.10	43.88	33.09	22.43	18.81	21.55	17.18	12.74	67
	mbest	74.16	43.96	32.67	21.98	18.37	20.76	16.44	12.07	
	mfinal	**75.60**	44.89	31.77	20.71	16.76	20.09	14.46	10.87	
Ours-W-RS	*best*	70.82	45.25	37.03	28.12	25.36	24.12	24.77	17.68	75
	mbest	70.66	45.51	36.50	27.51	24.75	23.97	23.38	17.14	
	mfinal	70.71	45.56	36.01	26.92	25.55	23.94	22.65	16.71	
Ours-E-RS	*best*	62.53	43.81	37.56	**31.59**	**29.68**	24.48	**28.11**	17.97	75
	mbest	62.38	42.07	36.78	30.80	28.90	23.64	27.71	17.55	
	mfinal	**77.20**	**47.74**	35.76	27.14	22.24	16.16	18.45	6.76	
Ours-B-RS	*best*	65.28	46.12	**38.11**	30.48	28.37	**26.72**	26.98	**19.82**	75
	mbest	65.42	45.94	37.54	30.01	27.85	26.28	26.52	19.43	
	mfinal	67.10	47.38	37.26	29.44	27.06	26.19	25.74	18.77	
NuAT [9]	*best*	74.25	45.01	35.45	26.05	23.78	24.17	22.48	18.53	101
	mbest	74.62	44.92	35.22	25.93	23.67	24.07	22.37	18.43	
	mfinal	75.29	45.31	34.85	25.58	23.44	23.62	22.10	18.06	
ATAS* [7]	-	64.11	-	31.39	-	28.15	-	-	21.09	-
FGSM-MEP [11]	*best*	55.29	37.14	32.42	27.40	26.61	22.39	26.05	19.01	92
	mbest	53.32	36.24	31.85	27.28	26.56	22.10	26.08	18.98	
	mfinal	**86.50**	**79.23**	10.09	0.06	0.04	0.02	2.42	0.06	
Ours-E-MEP	*best*	69.23	46.54	**41.21**	**34.66**	33.13	23.69	31.47	18.97	101
	mbest	69.84	47.18	40.90	34.17	32.72	22.69	31.12	17.74	
	mfinal	71.00	47.71	40.36	33.73	32.58	20.82	30.54	15.21	
Ours-B-MEP	*best*	63.30	45.25	40.31	34.50	**33.39**	**28.32**	**32.57**	**24.39**	101
	mbest	63.84	45.48	40.13	34.21	32.95	28.19	32.04	23.68	
	mfinal	64.69	46.07	39.95	33.84	32.39	27.79	31.42	22.55	

Table 9. Quantitative results of the adversarial training methods (ξ = 16/255) on CIFAR-10 with ResNet18 as the backbone. 'ATAS*' is the result of ATAS in [7], which is superior to our reproduction. We train each method three times. *mbest* (or *mfinal*) represents the evaluation average between the best (or final) models of three training processes. *best* is the best evaluation result for each FAT method. Weight centralization and regularization in MEP do not work together.

Methods		Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD↑	AA↑	Time (min)↓
PGD-AT	*best*	40.88	24.79	21.54	18.06	17.54	15.08	17.10	12.80	370
	mbest	40.57	24.72	21.46	17.94	17.38	14.98	17.02	12.63	
	mfinal	41.03	24.83	20.92	17.29	16.72	14.93	16.33	12.31	
FGSM-RS [10]	*best*	31.25	15.76	13.06	10.51	10.03	8.78	9.78	7.28	67
	mbest	30.12	15.24	12.69	10.25	9.79	8.45	9.57	6.90	
	mfinal	**52.36**	41.04	0.00	0.00	0.00	0.00	0.00	0.00	
GradAlign [1]	*best*	31.29	16.07	13.09	10.62	9.98	8.46	9.76	6.97	135
	mbest	31.91	15.71	12.56	10.28	9.71	8.22	9.47	6.61	
	mfinal	41.45	**43.36**	0.00	0.00	0.00	0.00	0.00	0.00	
ZeroGrad [5]	*best*	46.67	23.48	17.66	13.03	11.99	11.59	11.03	7.96	67
	mbest	47.31	23.58	17.60	12.85	11.87	11.62	11.01	7.94	
	mfinal	49.46	25.31	16.32	11.56	10.36	10.75	9.41	7.21	
Ours-W-RS	*best*	45.01	24.97	18.82	14.66	13.64	12.20	12.96	9.19	76
	mbest	44.68	25.19	18.75	14.50	13.53	12.18	13.80	9.00	
	mfinal	41.97	26.30	19.09	14.77	13.72	12.82	13.36	9.67	
Ours-E-RS	*best*	40.28	22.45	18.89	15.25	14.57	12.24	14.04	9.58	76
	mbest	41.09	22.33	18.78	15.19	14.43	12.00	13.91	9.50	
	mfinal	44.23	22.16	17.52	13.90	13.04	10.83	12.60	8.35	
Ours-B-RS	*best*	40.70	24.72	**19.61**	**15.76**	**14.75**	**13.03**	**14.29**	**10.25**	76
	mbest	41.47	25.98	19.44	15.36	14.34	12.91	13.71	9.90	
	mfinal	41.97	26.30	19.09	14.77	13.72	12.82	13.36	9.67	
NuAT [9]	*best*	34.63	22.34	17.06	14.55	13.91	11.73	12.59	8.44	101
	mbest	31.42	20.39	16.15	13.87	13.29	11.12	12.25	8.32	
	mfinal	43.73	27.11	14.54	10.40	8.83	9.32	7.30	4.97	
ATAS [7]	*best*	55.36	30.95	15.33	10.47	8.62	11.2	6.34	5.10	70
	mbest	55.63	30.35	15.31	10.28	8.49	10.97	6.30	5.05	
	mfinal	57.89	30.37	14.03	9.29	7.57	10.21	5.33	4.34	
FGSM-MEP [11]	*best*	21.29	13.32	11.46	9.97	9.76	7.32	9.58	6.29	92
	mbest	21.39	13.00	11.37	9.93	9.76	7.28	9.58	6.36	
	mfinal	**67.14**	**59.33**	1.43	0.64	0.32	0.39	0.17	0.01	
Ours-E-MEP	*best*	44.09	24.73	20.85	17.31	16.59	13.63	16.21	11.19	102
	mbest	44.00	24.46	20.59	17.11	16.59	13.37	16.05	10.97	
	mfinal	46.03	24.45	20.08	16.62	15.83	12.98	15.41	10.85	
Ours-B-MEP	*best*	41.86	24.71	**20.97**	**17.34**	**16.60**	**14.08**	**16.32**	**11.49**	102
	mbest	41.86	24.86	20.84	17.30	16.59	13.96	16.25	11.38	
	mfinal	43.10	24.33	20.60	17.13	16.50	13.68	16.16	11.30	

Table 10. Quantitative results of the adversarial training methods (ξ = 16/255) with ResNet18 as the backbone on CIFAR-100. We train each method three times. *mbest* (or *mfinal*) represents the evaluation average of the best (or final) model in three training processes. *best* is the best evaluation result for each FAT method.

Methods		Clean↑	FGSM↑	PGD-10↑	PGD-20↑	PGD-50↑	C&W↑	APGD-CE↑	AA↑	Time (hour)↓
PGD	*best*	32.47	16.37	13.27	10.60	10.23	8.05	10.01	6.41	67.2
	mbest	32.52	16.47	13.40	10.63	10.24	7.95	10.00	6.41	
	mfinal	32.32	16.26	13.35	10.25	9.83	7.56	9.69	6.21	
FGSM-RS [10]	*best*	31.25	13.66	10.24	7.01	6.33	5.25	6.00	3.76	10.5
	mbest	27.48	12.46	9.59	6.97	6.47	4.99	6.19	3.63	
	mfinal	0.00	2.39	0.00	0.00	0.00	0.00	0.00	0.00	
GradAlign [1]	*best*	29.15	13.74	10.60	7.79	7.13	5.89	6.71	4.08	20.9
	mbest	28.65	13.80	10.40	7.78	7.08	5.75	6.59	3.90	
	mfinal	15.66	8.72	5.75	4.55	4.38	3.05	4.25	2.24	
ZeroGrad [5]	*best*	35.13	12.21	8.42	5.43	4.77	3.88	4.34	2.36	10.5
	mbest	34.66	12.26	8.22	5.29	4.82	3.71	4.23	2.21	
	mfinal	37.67	7.62	3.18	1.70	1.38	1.07	1.05	0.51	
NuAT [9]	*best*	34.55	15.38	12.18	8.96	8.43	6.38	8.19	4.33	24.6
	mbest	35.24	15.52	12.07	8.81	8.13	6.29	7.68	4.27	
	mfinal	41.75	16.51	10.49	7.01	6.37	4.93	5.35	2.94	
FGSM-BP [11]	*best*	21.44	10.39	8.62	7.11	6.91	5.00	6.77	3.80	14.3
	mbest	20.02	10.18	8.38	6.98	6.77	4.56	6.57	3.62	
	mfinal	**49.56**	**38.54**	0.00	0.00	0.00	0.00	0.00	0.00	
B-BP (Ours)	*best*	34.70	16.55	**13.43**	**10.49**	**10.07**	**7.82**	**9.76**	**6.28**	15.4
	mbest	33.51	16.32	12.92	9.72	9.23	7.26	9.32	5.95	
	mfinal	33.28	15.81	12.47	8.69	8.25	6.61	8.18	5.67	

Table 11. Quantitative results of various methods (ξ = 16/255) with PreActResNet18 as the backbone on Tiny ImageNet. *best* is the best evaluation result among the three training sessions for each FAT method.