ELIMINATING CATASTROPHIC OVERFITTING VIA ABNORMAL ADVERSARIAL EXAMPLES REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-step adversarial training (SSAT) is shown to be able to defend against iterative-step adversarial attacks to achieve both efficiency and robustness. However, SSAT suffers from catastrophic overfitting (CO) with strong adversaries, showing that the classifier decision boundaries are highly distorted and robust accuracy against iterative-step adversarial attacks suddenly drops from peak to nearly 0% in a few epochs. In this work, we find that some adversarial examples generated on the network trained by SSAT exhibit anomalous behaviour, that is, although the training data is generated by the inner maximization process, the loss of some adversarial examples decreases instead, which we called abnormal adversarial examples. Furthermore, network optimization on these abnormal adversarial examples will further accelerate the model decision boundaries distortion, and correspondingly, the number of abnormal adversarial examples will sharply increase with CO. These observations motivate us to eliminate CO by hindering the generation of abnormal adversarial examples. Specifically, we design a novel method, *Abnormal Adversarial Examples Regularization* (AAER), which explicitly regularizes the number and outputs variation of abnormal adversarial examples to hinder the model from generating abnormal adversarial examples. Extensive experiments demonstrate that our method can eliminate CO and further boost adversarial robustness with strong adversaries.

1 INTRODUCTION

In recent years, Deep Neural Networks (DNNs) have performed impressively in various fields, such as autonomous driving (Litman, 2017), face recognition (Sharif et al., 2016) and medical imaging diagnosis (Buch et al., 2018). However, DNNs were found to be vulnerable to adversarial examples (Szegedy et al., 2013). Although these adversarial examples are imperceptible to the human eyes, they can lead to a completely different prediction in DNNs. To this end, many adversarial defense methods have been proposed, such as verification and provable defense (Katz et al., 2017), pre-processing techniques (Guo et al., 2017), detection algorithms (Metzen et al., 2017) and adversarial training (AT) (Goodfellow et al., 2014). Among them, AT is considered to be one of the most effective methods against adversarial attacks (Athalye et al., 2018). However, standard iterative-step AT significantly increases computational overhead due to multiple steps forward and backward propagation.

Therefore, some works attempt to improve the vanilla single-step adversarial training (SSAT) to defend against iterative-step adversarial attacks while maintaining efficiency and robustness. Unfortunately, a serious problem - catastrophic overfitting (CO) - occurs with stronger adversaries (Wong et al., 2020). This strange phenomenon means that the robust accuracy of the model against the iterative-step adversarial attack suddenly from peak drops to nearly zero during a few epochs, as shown in Figure 1. This intriguing phenomenon has been widely investigated and led to many works to resolve CO. Recently, Kim et al. (2021) points out that networks in which CO occurs are generally accompanied by highly distorted decision boundaries. However, the interaction between distorted decision boundaries and CO remains unclear. In this work, we delve into the dynamic effects between CO and decision boundaries distortion. Specifically, we find some adversarial examples generated on the network with distorted decision boundaries exhibit anomalous behavior,

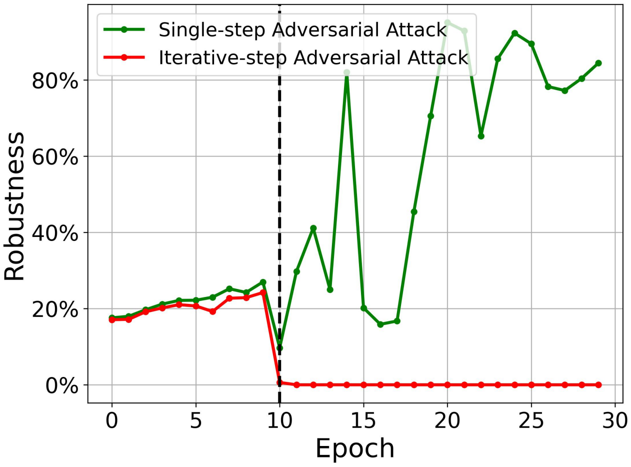

Figure 1: Model robust test accuracy with different noise magnitudes. The red and green lines are defence against FGSM and PGD-7-1 adversarial attack, respectively. The dashed line and solid line noise magnitude are 8/255 and 16/255, respectively. Dashed black lines correspond to the 10th epoch, which is the point that model occurs CO.

that is, although all training samples are generated by the inner maximization process, the loss of some adversarial examples decreases instead. We refer to these training samples as abnormal adversarial examples. To make matters worse, the decision boundaries distortion will further exacerbate by optimising the classifier directly on these abnormal adversarial examples, and the number of abnormal adversarial examples will sharply increase as a result, which leads to a vicious circle between the number of abnormal adversarial example and the decision boundaries distortion. All these atypical findings raise a question:

Can CO be prevented by hindering the generation of abnormal adversarial examples?

To answer the above question, we design a novel method, *Abnormal Adversarial Examples Regularization* (AAER), which incorporates a regularizer that prevents CO by suppressing generated abnormal adversarial examples. Specifically, AAER consists of two key components: (i) the number and (ii) outputs variation of abnormal adversarial examples. The first part (i) counts the sample number by dividing the training samples into groups of normal and abnormal adversarial examples through anomalous loss decrease behavior. The second part (ii) contains prediction confidence and logits variation, and calculates these two variations differences between the two groups of samples by cross-entropy and Euclidean distance, respectively. Then, AAER explicitly regularizes the number and outputs variation of abnormal adversarial examples by these two parts to hinder the model from generating abnormal adversarial examples. Extensive experiments show that our method can well eliminate CO and further improve the adversarial robustness. It is worth noting that our method does not involve the extra generation and backward propagation process, which brings us unparalleled convenience in computational overhead. Our major contributions are summarized as follows:

- We found some training samples exhibit anomalous loss variation during the inner maximization process. Besides, the number of abnormal adversarial examples will sharply increase with CO, and the model will further exacerbate by optimising the classifier directly on these abnormal adversarial examples.
- Based on the observed effect, we propose a novel method - *Abnormal Adversarial Examples Regularization* (AAER), which explicitly regularizes the number of abnormal adversarial examples and their anomalous outputs variation to hinder the generation of abnormal adversarial examples. Extensive experiments demonstrate that our method can prevent CO and automatically adapt to different noise magnitudes without hyperparameter tuning.
- We evaluate the effectiveness of our method across different adversarial budgets, adversarial attacks, datasets and network architectures, showing that our proposed method consistently achieves state-of-the-art robust accuracy in SSAT and can obtain comparable robustness to standard iterative-step AT with only negligible computational overhead.

2 Related Work

2.1 Adversarial Training

Adversarial training has been demonstrated to be the most effective method for defending against adversarial attacks (Athalye et al., 2018). AT is generally formulated as a min-max optimization problem (Madry et al., 2017), the inner maximization problem tries to generate the strongest adversarial examples to maximise the loss, and the outer minimization problem tries to optimize the network to minimize the loss on adversarial examples. However, the inner maximization problem is a NP-hard problem. Therefore, AT uses a simple gradient ascent to generate perturbations to find local approximate solution, and can be formalized as a min-max optimization problem as follows:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\Delta} \ell(x+\delta, y; \theta) \right], \tag{1}$$

where (x, y) is the training dataset from the distribution D, $\ell(x, y; \theta)$ is the loss function parameterized by θ, δ is the perturbation confined within the boundary ϵ with L_p- norm distance, shown as: $\Delta = \{\delta : \|\delta\|_p \leq \epsilon\}$. The common threat models are L_1, L_2 and L_∞, in this work we chose L_∞ for our threat model.

Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) is a single-step adversarial attack method, which uses the sign of the gradient to find the perturbation, as shown in Eq. (2):

$$\delta_{FGSM} = \epsilon \cdot \text{sign}\left(\nabla_x \ell(x, y; \theta)\right). \tag{2}$$

Fast Training (RS-FGSM) (Wong et al., 2020) adds uniform random initialization η before generating the perturbation, and uses the over-perturbation step size $\alpha = 1.25 \cdot \epsilon$:

$$\begin{aligned} \eta &= \text{Uniform}(-\epsilon, \epsilon), \\ \delta_{RS-FGSM} &= \alpha \cdot \text{sign}\left(\nabla_{x+\eta} \ell(x+\eta, y; \theta)\right). \end{aligned} \tag{3}$$

Iterative Fast Gradient Sign Method (I-FGSM) (Kurakin et al., 2018) is an iterative-step version of FGSM that uses multiple gradients to find stronger perturbations. With a smaller step size $\alpha = \epsilon/N$ and the number of iterations T, I-FGSM can be formulated as follows:

$$\delta^T_{I-FGSM} = \alpha \cdot \text{sign}\left(\nabla_{x+\delta^{T-1}} \ell(x+\delta^{T-1}, y; \theta)\right). \tag{4}$$

Projected Gradient Descent (PGD) (Madry et al., 2017) adds uniform random initialization on the basis of I-FGSM.

2.2 Catastrophic Overfitting

Since Wong et al. (2020) found the CO phenomenon, there has been a line of work trying to explore and mitigate this problem. Vivek & Babu (2020b) empirically showed that adding a dropout layer after all non-linear layers can avoid early overfitting to FGSM. de Jorge et al. (2022) found that augmenting the perturbations by increasing the noise initialization magnitude and removing the perturbation boundaries can eliminate CO. Li et al. (2022) successfully prevents CO by constraining training samples in a carefully extracted subspace to avoid the abrupt growth of gradient.

Other works attempt to prevent CO by strengthening the inner maximization processes. Kim et al. (2021) assumed that CO is caused by fixed FGSM perturbation magnitude and reduces the perturbed step size for misclassified adversarial examples. Golgooni et al. (2021) argued that small gradients play a key role in CO and ignore small gradient information to avoid huge weight updates. Huang et al. (2022) discovered that fitting instances with a larger gradient norm are more likely to cause CO and learning an instance-adaptive step size is inversely proportional to its gradient norm. Park & Lee (2021) leverages the gradients of latent representation as the latent adversarial perturbation to compensate for local linearity.

Similar to our work, some works add a regularization term on the loss value to explicitly prevent CO. Andriushchenko & Flammarion (2020) found that PGD and FGSM perturbations are orthogonal when CO occurs, hence they proposed a regularization term to encourage the gradient alignment. Vivek & Babu (2020a) proposed a regularization term to mitigate the CO by harnessing properties that differentiate a robust model from that of a pseudo-robust model. Sriramanan et al. (2021) introduced a relaxation term to find more suitable gradient directions for the attack by smoothing the loss

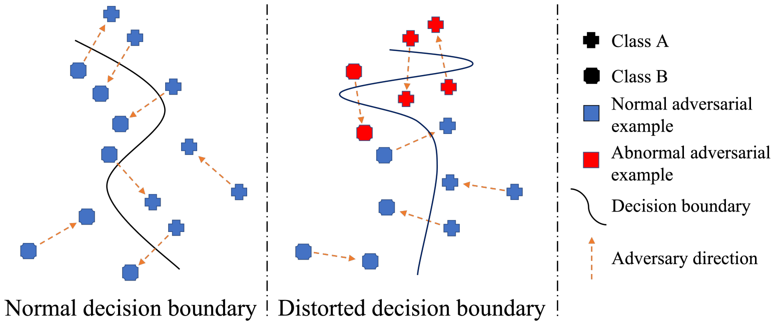
Class A
Class B
Normal adversarial example
Abnormal adversarial example
Decision boundary
Adversary direction
Normal decision boundary
Distorted decision boundary

Figure 2: Visualization of classifier decision boundary and training samples. The left panel shows that the training samples generated on the normal decision boundary are all belong to the normal adversarial example (blue) which can mislead the classifier. The middle panel shows that some training samples generate on the distorted decision boundary that cannot mislead the classifier, which we called abnormal adversarial example (red).

surface. Chen et al. (2021) demonstrated that the negative high-order terms lead to a perturbation loss distortion phenomenon that will cause CO, and they proposed a regularization term to make the loss surface flat.

3 Proposed Approach

In this section, we first define abnormal adversarial example and show how their number change during CO (Section 3.1). We further analyse the outputs variation of normal and abnormal adversarial examples and find that they exhibit significantly different magnitudes of outputs variation after CO (Section 3.2). Based on our observations, we propose a novel regularization term, AAER, using the number and outputs variation of abnormal adversarial examples to explicitly suppress the generation of these anomalous training samples to eliminate CO (Section 3.3).

3.1 Definition and Counting the Abnormal Adversarial Example

Adversarial training employs the most adversarial data to reduce the sensitivity of the network's output w.r.t. adversarial perturbation of the natural data. Therefore, we expect the inner maximization process can generate effective adversarial examples that can maximize the classification loss. However, Kim et al. (2021) shows that the decision boundaries of the classifier will be highly distorted accompanied by the occurrence of CO. After adding the adversarial perturbation which is generated on this distorted classifier, the classification loss of some training samples is atypically reduced. As shown in Figure 2, it can be seen that, for some samples (blue), they will misclassify the model or be closer to the decision boundary after the inner maximization process, and for some other samples (red), they are farther to the decision boundary after adding the perturbation generated by the distorted classifier, which we called abnormal adversarial example. These abnormal adversarial examples generally fail to mislead the classifier. Thus, we can define abnormal adversarial examples using the following formula:

$$\begin{aligned} \delta &= \alpha \cdot \operatorname{sign}\left(\nabla_{x+\eta} \ell(x+\eta, y ; \theta)\right), \\ \delta^{Abnormal} &\stackrel{def}{=} \ell\left(x+\eta, y ; \theta\right)>\ell\left(x+\eta+\delta, y ; \theta\right). \end{aligned} \tag{5}$$

We further observe the changes in the number of abnormal adversarial examples during model training, and the statistical results are shown in Figure 3 (left). It can be observed that before CO occurs, the number of abnormal adversarial examples is very small, almost close to 0. During the occurrence of CO, their number increases sharply. For example, the number of abnormal adversarial examples surged 39 times (red line) at the 10th epoch. Note that the rapid increase in the number of abnormal adversarial examples also implies that the classifier boundaries are continuously deteriorating, which also leads to a further rise in the number of abnormal adversarial examples and peaks at the 13th epoch, which is approximately 55 times the number of abnormal adversarial examples before CO occurred. After the occurrence of CO, the abnormal adversarial examples are basically maintained

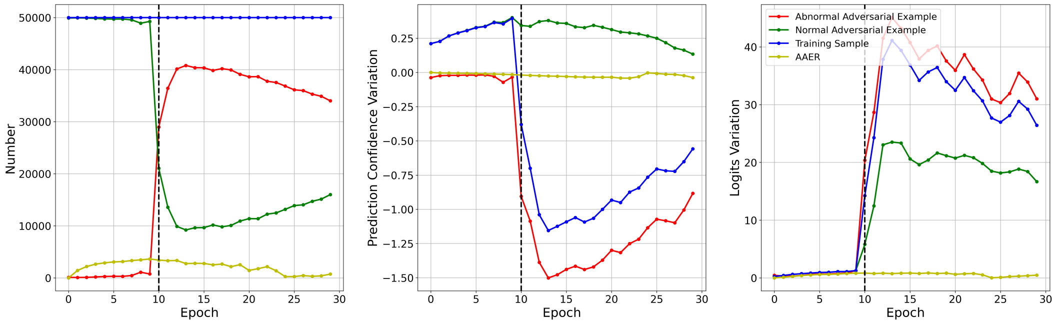

Figure 3: The number, prediction confidence and logits variation of normal/abnormal adversarial examples and training samples. The left, middle and right panel shows the number, prediction confidence and logits variation, respectively. The green/red and blue lines represent normal/abnormal adversarial examples and training samples, respectively. Dashed black lines correspond to the 10th epoch, which is the point that model occurs CO. The yellow line represents the number, prediction confidence and the logits variation of abnormal adversarial examples under the AAER method.

at a very large number. Given this observation, we can infer that there is a close correlation between the number of abnormal adversarial examples and the CO phenomenon, which also prompts us to wonder (Q1): *whether CO can be mitigated by reducing the number of abnormal adversarial examples.*

3.2 Outputs Variation of Normal and Abnormal Adversarial Example

The above observations indicate that CO and the number of abnormal adversarial examples are closely related. In this part, we further analyze the outputs variation of normal and abnormal adversarial examples during CO. Specifically, we analyze the outputs into two categories: prediction confidence and logits, and use the cross-entropy to calculate the prediction confidence variation during the inner maximization process, which is formulated as follows:

$$\ell\left(x+\eta+\delta, y ; \theta\right)-\ell\left(x+\eta, y ; \theta\right). \tag{6}$$

From Figure 3 (middle), we can observe that the change in the prediction confidence of normal adversarial samples is greater than 0, indicating that the perturbation leads to worse predictions. However, the variation of abnormal adversarial examples is atypical negative, meaning that the perturbation has the opposite effect as we expected. Furthermore, we analyze the prediction confidence variation of abnormal adversarial examples during training. Before the occurrence of CO, we can observe that the prediction confidence variation of abnormal adversarial examples is slightly less than zero, and the negative impact on all training samples (blue line) is not significant. During the occurrence of CO, their prediction confidence variation decreases rapidly slumped 27 times at the 10th epoch, and deterministically effect the prediction confidence of all training samples.

Furthermore, we compare the magnitude of the outputs change between normal and abnormal adversarial examples, and use the Euclidean distance (L2 distance) to calculate the sample logits variation during the inner maximization process, which is formulated as follows:

$$\| f_{\theta}\left(x+\eta+\delta\right)-f_{\theta}\left(x+\eta\right) \|_{2}^{2}, \tag{7}$$

where f_θ is the DNN classifier parameterized by θ and $\|\cdot\|_2^2$ is the L2 distance.

The magnitude of the logits variation of normal and abnormal adversarial examples is shown in Figure 3 (right). We can observe that the logits variation magnitude of abnormal adversarial examples increases dramatically during CO, which is 16 times larger than that before CO. A single-step gradient ascent can bring an earth-shaking change in the output logits, which generally happens on highly distorted decision boundaries. Additionally, we observed that the logits variation magnitude of the normal adversarial examples (green line) increases one epoch later than the abnormal ones, which indicates that the model boundary distortion mainly lies in the abnormal adversarial examples, in other words, directly optimizing the network with these abnormal adversarial examples will further exacerbate the model boundary distortion. Moreover, we further compare the magnitude of logits variation for normal and abnormal adversarial examples. From Figure 3 (right), we can observe that the logits variation magnitude on normal and abnormal adversarial examples is similar before CO. However, there is a significant difference in the logits variation magnitude between these two types of examples after CO. It is observed that the logits variation magnitude in abnormal adversarial examples is 4 times that in normal ones at the 10th epoch. There are significant differences in the magnitude of both prediction confidence and logits variation between normal and abnormal adversarial examples, which inspires us to wonder (Q2): *whether CO can be mitigated by constraining the outputs variation of abnormal adversarial examples.*

3.3 Abnormal Adversarial Examples Regularization Term

We answer these two questions through three optimization objectives. To answer the Q1, the first part (i) uses the Eq. 5 to divide the training samples into normal and abnormal adversarial examples, and then penalize the number of abnormal adversarial examples. To answer the Q2, the second part (ii) and the third part (iii) constrain the outputs variation of abnormal adversarial examples. Specifically, the second part (ii) calculates the prediction confidence variation of abnormal adversarial examples, and then penalizes this variation that should not decrease during the inner maximization process, which is formalized as follows:

$$\frac{1}{n}\sum_{j=1}^{n}\left(\ell\left(x_j^{Abnormal}+\eta, y_j;\theta\right)-\ell\left(x_j^{Abnormal}+\eta+\delta, y_j;\theta\right)\right), \tag{8}$$

where n is the number of abnormal adversarial examples.

The third part (iii) calculates the logits variation of normal and abnormal adversarial examples. Since the logits variation is a representation of the change magnitude, which is not related to the label, there is no clear target value for the optimization standard. Therefore, we use the logits variation of normal adversarial examples as the standard and explicitly make them logits variation closer. In order to avoid the network only focusing on increasing the logits variation of abnormal adversarial examples instead of reducing the abnormal ones, we use the max function to limit the minimum value to 0, which is formalized as follows:

$$\begin{aligned} max\Bigg(&\frac{1}{n}\sum_{j=1}^{n}\left(\|f_\theta\left(x_j^{Abnormal}+\eta+\delta\right)-f_\theta\left(x_j^{Abnormal}+\eta\right)\|_2^2\right) \\ &-\frac{1}{m-n}\sum_{k=1}^{m-n}\left(\|f_\theta\left(x_k^{Normal}+\eta+\delta\right)-f_\theta\left(x_k^{Normal}+\eta\right)\|_2^2\right), 0\Bigg), \end{aligned} \tag{9}$$

where m is the number of training samples and $max(,)$ is the max function.

Based on the above analysis, we design a novel regularization term, AAER, which aims to suppress the abnormal adversarial examples by (i) the number, (ii) the prediction confidence variation and (iii) the logits variation, ultimately achieving the purpose of preventing CO, which is shown in the following formula:

$$AAER=\frac{n}{m}\cdot\left(Eq.\,8\cdot\lambda 1+Eq.\,9\cdot\lambda 2\right), \tag{10}$$

where $\lambda 1$ and $\lambda 2$ is the hyperparameter to control the strength of the regularization term.

AAER can effectively hinder the generation of abnormal adversarial examples which are highly correlated with distorted classifier, thereby encouraging training of smoother classifiers that can better defend against adversarial attacks. Furthermore, the strength of AAER depends on the product of the number and outputs variation of abnormal adversarial examples, which can more comprehensively and flexibly reflect the degree of classifier distortion. The algorithm realization is summarized in Algorithm 1. Note that we employ increasing α to stabilize the optimization objective and avoid model training to crash in the early stages.

4 Experiment

In this section, we conduct extensive experiments to verify the effectiveness of AAER including experiment settings (Section 4.1), performance evaluations (Section 4.2), ablation studies (Section 4.3) and time complexity study (Section 4.4).

Algorithm 1 *Abnormal Adversarial Examples Regularization* (AAER)

Input: network f_θ, epochs T, mini-batch M, perturbation radius ϵ, step size α, initialization term η.

```
for $t = 1 \ldots T$ to do
    for $i = 1 \ldots M$ to do
        $\alpha = t/T \cdot \alpha$
        $\delta = \alpha \cdot \operatorname{sign}\left(\nabla_{x+\eta} \ell(x_i + \eta, y_i; \theta)\right)$
        $CEloss = \frac{1}{m} \sum_{i=1}^{m} \ell\left(x_i + \eta + \delta, y_i; \theta\right)$
        $AAERloss$ = Eq. (10)
        $\theta = \theta - \nabla_\theta \left(CEloss + t/T \cdot AAERloss, \right)$
    end for
end for
```

4.1 Experiment Setting

Baselines. We compare our method with other SSAT methods including RS-FGSM (Wong et al., 2020), ATTA (Zheng et al., 2020), FreeAT (Shafahi et al., 2019), N-FGSM (de Jorge et al., 2022), Grad Align (Andriushchenko & Flammarion, 2020), ZeroGrad and MultiGrad (Golgooni et al., 2021). We also compare our method with iterative-step AT PGD-2 and PGD-10 (Madry et al., 2017) providing a reference for the ideal performance. To accommodate different adversarial budgets, we use PGD-10 with two step size of 2/255 and ϵ/10. We will show natural and robust accuracy results using the hyperparameters reported in their official repository (except for FreeAT, we do not divide the number of epochs by m to keep the same training epochs). It is worth noting that we do not use early stopping (Wong et al., 2020) as this technique can restore the robustness of all methods.

Datasets and Model Architectures. We will show the results on the benchmark datasets Cifar-10/100 (Krizhevsky et al., 2009) and use random cropping and horizontal flipping for data argumentation. We use the PreactResNet-18 (He et al., 2016) and WideResNet-34 (Zagoruyko & Komodakis, 2016) architectures on these datasets to evaluate results. The training results of WideResNet-34 are also available in the Appendix B. We also report the settings and results of our method on SVHN (Netzer et al., 2011) and Tiny-imagenet (Netzer et al., 2011) in the Appendix E.

Attack Methods and Learning Rate Schedule. To report the robust accuracy of models, we attack these methods using the standard PGD adversarial attack with $\alpha = \epsilon/4$, 10 restarts and 50 attack steps. We also evaluate our methods based on Auto Attack in the Appendix C. We use the cyclical learning rate schedule (Smith, 2017) with 30 epochs that reaches its maximum learning rate (0.2 in our experiments) when half of the epochs (15) are passed on Cifar-10/100.

Setup for Our Proposed Method. In this work, we use the SGD optimizer with momentum of 0.9 and weight decay of 5 × 10^{-4}. We chose L_∞ as the threat model and set gradient ascent step size $\alpha = 1.5 \cdot \epsilon$. We set η = Uniform$(-\epsilon, \epsilon)$ for random initialization, and the η setting for previous initialization can be found in the Appendix A. We will show the best λ settings in the Appendix D. It is worth noting that our method can also achieve robustness without tuning hyperparameters with different adversarial budgets, the results on universal λ in the Appendix D.

4.2 Performance Evaluation

In this part, we report the experimental results of our method under four different settings AAER$_{\mathrm{RC}}$: AAER with random initialization and clipped perturbations and AAER$_{\mathrm{RUC}}$: AAER with random initialization and unclipped perturbations. The unclipped technique was proposed by de Jorge et al. (2022), who claimed that clipping is performed after taking a gradient ascent step, which may make adversarial samples no longer effective. The AAER based on the previous initialization is available in the Appendix A.

CIFAR10 Results. In Table 1, we present an evaluation of the proposed methods with the competing baselines on the CIFAR-10 dataset. First, we can observe that RS-FGSM, ATTA and FreeAT suffer from CO with strong adversaries. We can also observe an interesting phenomenon that some weakly robust methods will recover partial robustness with large noise magnitude 32/255. Table 1 shows that our proposed methods can significantly improve the robust accuracy, achieve superior

Table 1: CIFAR10/100: Accuracy of different methods and different noise magnitudes using PreActResNet-18 under L_∞ threat model. The left and right panel are the CIFAR10 and CIFAR100 results, respectively. The top number is the natural accuracy while the bottom number is the PGD-50-10 accuracy. The results are averaged over 3 random seeds and reported with the standard deviation.

dataset	CIFAR10				CIFAR100			
noise magnitude	8/255	12/255	16/255	32/255	8/255	12/255	16/255	32/255
RS-FGSM	83.91 ± 0.21	66.46 ± 22.80	66.54 ± 12.25	36.43 ± 7.86	60.29 ± 1.51	18.19 ± 8.51	11.03 ± 5.24	11.40 ± 8.60
	46.01 ± 0.18	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	10.58 ± 13.10	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
ATTA	86.41 ± 0.39	81.15 ± 0.94	82.75 ± 0.71	39.27 ± 4.48	61.25 ± 0.25	37.40 ± 16.34	47.14 ± 11.12	27.71 ± 6.93
	42.15 ± 0.42	19.55 ± 15.20	0.00 ± 0.00	7.84 ± 6.80	22.78 ± 0.19	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
FreeAT	76.20 ± 1.09	68.07 ± 0.38	45.84 ± 19.07	61.11 ± 8.41	47.41 ± 0.30	39.84 ± 0.40	3.32 ± 2.48	26.23 ± 15.54
	43.74 ± 0.41	33.14 ± 0.62	0.00 ± 0.00	0.00 ± 0.00	22.27 ± 0.33	16.57 ± 0.20	0.00 ± 0.00	0.00 ± 0.00
ZeroGrad	81.60 ± 0.16	77.52 ± 0.21	79.65 ± 0.17	65.48 ± 6.26	53.83 ± 0.22	49.07 ± 0.14	50.76 ± 0.02	47.75 ± 2.99
	47.56 ± 0.16	27.34 ± 0.09	6.37 ± 0.23	0.00 ± 0.00	25.02 ± 0.24	14.76 ± 0.26	5.23 ± 0.09	0.01 ± 0.02
MultiGrad	81.65 ± 0.16	81.09 ± 4.67	82.98 ± 3.30	41.90 ± 30.53	53.11 ± 0.34	46.81 ± 0.51	46.05 ± 8.68	16.25 ± 10.48
	47.93 ± 0.18	9.95 ± 16.97	0.00 ± 0.00	5.69 ± 5.14	25.68 ± 0.21	16.56 ± 0.56	0.00 ± 0.00	0.53 ± 0.91
N-FGSM	80.48 ± 0.21	71.30 ± 0.12	62.96 ± 0.74	32.44 ± 2.79	54.87 ± 0.28	46.16 ± 0.13	37.76 ± 0.16	15.56 ± 4.55
	47.91 ± 0.29	36.23 ± 0.10	27.14 ± 1.44	10.58 ± 0.89	26.51 ± 0.38	18.75 ± 0.19	**14.14 ± 0.05**	1.61 ± 2.77
Grad Align	82.10 ± 0.78	74.17 ± 0.55	60.37 ± 0.95	25.23 ± 3.41	54.00 ± 0.44	45.83 ± 0.72	36.80 ± 0.10	15.05 ± 0.07
	47.77 ± 0.58	34.87 ± 1.00	27.90 ± 1.01	11.53 ± 3.23	25.27 ± 0.68	18.13 ± 0.71	13.77 ± 0.76	2.85 ± 1.34
AAER_{RC}	84.33 ± 0.08	76.07 ± 0.78	65.88 ± 0.60	26.14 ± 1.05	58.94 ± 0.53	49.09 ± 0.16	39.44 ± 0.88	20.03 ± 1.08
	46.42 ± 0.19	32.94 ± 0.31	24.67 ± 0.52	**14.70 ± 0.39**	25.34 ± 0.21	16.94 ± 0.24	12.33 ± 0.14	**5.20 ± 0.17**
AAER_{RUC}	77.41 ± 0.13	66.24 ± 0.60	55.32 ± 0.55	20.57 ± 2.61	53.60 ± 0.19	40.56 ± 0.33	29.27 ± 0.51	12.47 ± 1.26
	51.51 ± 0.22	**38.66 ± 0.38**	**30.31 ± 0.15**	12.72 ± 0.52	**28.50 ± 0.15**	**19.69 ± 0.23**	13.96 ± 0.07	4.61 ± 0.09
PGD-2	85.07 ± 0.12	78.97 ± 0.23	72.31 ± 0.40	48.45 ± 0.71	60.09 ± 0.20	53.46 ± 0.27	47.50 ± 0.28	31.89 ± 0.69
	45.27 ± 0.07	32.99 ± 0.46	24.32 ± 0.64	11.24 ± 0.40	24.58 ± 0.12	17.16 ± 0.21	12.69 ± 0.06	4.51 ± 0.21
PGD-10 (2/255)	80.55 ± 0.37	72.37 ± 0.31	67.20 ± 0.69	59.35 ± 0.84	55.05 ± 0.25	47.42 ± 0.29	42.39 ± 0.17	34.68 ± 0.23
	50.67 ± 0.40	**38.60 ± 0.39**	**29.34 ± 0.18**	5.80 ± 0.23	**27.87 ± 0.12**	**20.29 ± 0.18**	**15.01 ± 0.21**	3.81 ± 0.12
PGD-10 (ϵ/10)	84.74 ± 0.11	78.31 ± 0.57	71.19 ± 0.46	40.51 ± 0.78	59.41 ± 0.37	52.74 ± 0.14	46.36 ± 0.32	26.07 ± 0.01
	46.06 ± 0.32	34.13 ± 0.62	26.07 ± 0.69	**15.16 ± 0.34**	24.71 ± 0.11	17.62 ± 0.13	13.39 ± 0.10	**6.46 ± 0.22**

robustness to other SSAT methods and even have comparable robustness to PGD AT. Interestingly, our unclipped perturbations methods always have better performance, except for 32/255 noise magnitude, where we conjecture that the unclipped perturbations are too large to disturb the original features of the inputs.

CIFAR100 Results. We also conduct experiments on the CIFAR100 dataset. Note that CIFAR100 is more challenging than CIFAR10 as the number of classes/training images per class is ten times larger/smaller than that of CIFAR10. As shown by the results in Table 1, the proposed methods are still able to prevent CO and improve robust accuracy. It verifies that the AAER can reliably prevent CO and is general across different datasets.

4.3 Ablation Study

In this part, we investigate the impacts of AAER_{RC} with 16/255 noise magnitude using PreactResNet-18 on CIFAR10 dataset under L_∞ threat model.

Optimization Objectives. To verify the effectiveness of our proposed method, we show the change in the three optimization objectives during training in Figure 3. We can observe that the number, prediction confidence and logits variation of abnormal adversarial examples are well constrained by AAER throughout the training. We also try to simply ignore abnormal adversarial examples and train only on normal ones. Unfortunately, this method does not work due to the abnormal

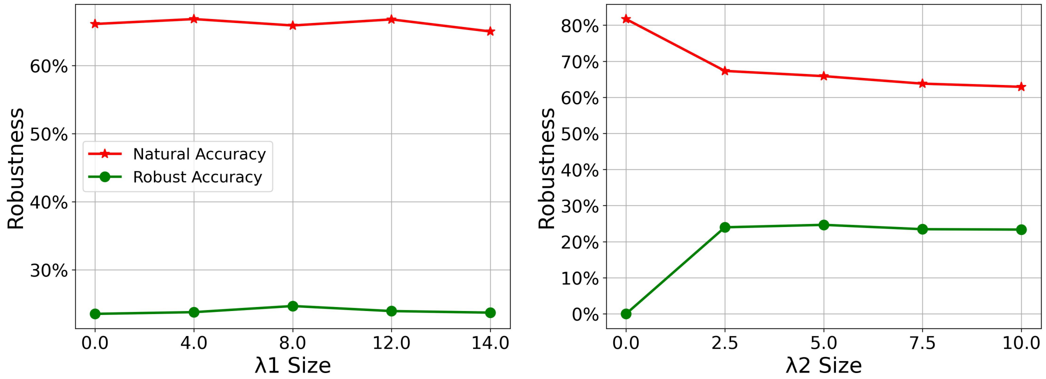

Figure 4: Ablation Study. The red and green line are the natural and robust test accuracy, respectively. **Left panel:** Effect of different sizes $\lambda 1$, and we fix $\lambda 2$ as 5.0 at this experiment. **Right panel:** Effect of different sizes $\lambda 2$, and we fix $\lambda 1$ as 8.0 at this experiment.

adversarial example is not the cause of decision boundary distortion but rather co-occurs. Thus, ignoring abnormal adversarial examples cannot repair existing decision boundary distortion.

λ **Selection.** To verify the effectiveness of our proposed method, we investigate the effect of different sizes $\lambda 1$ and $\lambda 2$ on natural and robustness performance. From the figure 4 (left), we can observe that the effect of $\lambda 1$ does not seem to be significant. However, it acts as a buffer to prevent the AAER from changing too drastically, and we chose $\lambda 1$ of 8.0 for optimal preference. From the Figure 4 (right), we can observe that the model can successfully prevent CO when $\lambda 2$ is not 0, which proves that our method can effectively eliminate CO. Under the same experimental setting mentioned before, with the value of λ varying from 0 to 10.0, we can observe that choosing $\lambda 2$ of 5.0 can achieve the best robustness.

4.4 Time Complexity Study

We show the time complexity of different AT methods in Table 2, we can observe that the running time for one epoch of AAER is almost equal to the RS-FGSM method. In contrast, Grad Align and PGD-10 are 2.3 and 4.6 times slower than our method, respectively.

Table 2: CIFAR10 training time on a single NVIDIA Tesla V100 GPU using PreactResNet-18. The results are averaged over 30 epochs.

Method	RS-FGSM	ATTA	FreeAT	ZeroGrad	MultiGrad
Second / Epoch	26.1S	41.5S	106.6S	28.7S	52.1S
Method	N-FGSM	Grad Align	AAER (Our)	PGD-2	PGD-10
Second / Epoch	25.9S	69.4S	30.5S	39.1S	140.7S

5 Conclusion

In this paper, we find that the abnormal adversarial examples exhibit anomalous behaviour, i.e. they are further to the decision boundaries after adding perturbations generated by the inner maximization process. We empirically show that the catastrophic overfitting is closely related to the abnormal adversarial examples by analyzing their number and outputs variation during model training. Motivated by this, we propose a novel and effective method, *Abnormal Adversarial Examples Regularization* (AAER), through a regularizer to eliminate catastrophic overfitting by suppressing generated abnormal adversarial examples. Our approach can successfully resolve the catastrophic overfitting with different noise magnitudes and achieve state-of-the-art preference with computational convenience in various settings.

REFERENCES

Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Varun H Buch, Irfan Ahmed, and Mahiben Maruthappu. Artificial intelligence in medicine: current trends and future possibilities. *British Journal of General Practice*, 68(668):143–144, 2018.

Renjie Chen, Yuan Luo, and Yisen Wang. Towards understanding catastrophic overfitting in fast adversarial training. 2021.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Pau de Jorge, Adel Bibi, Riccardo Volpi, Amartya Sanyal, Philip HS Torr, Grégory Rogez, and Puneet K Dokania. Make some noise: Reliable and efficient single-step adversarial training. *arXiv preprint arXiv:2202.01181*, 2022.

Zeinab Golgooni, Mehrdad Saberi, Masih Eskandar, and Mohammad Hossein Rohban. Zerograd: Mitigating and explaining catastrophic overfitting in fgsm adversarial training. *arXiv preprint arXiv:2103.15476*, 2021.

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

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

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.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International conference on computer aided verification*, pp. 97–117. Springer, 2017.

Hoki Kim, Woojin Lee, and Jaewook Lee. Understanding catastrophic overfitting in single-step adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8119–8127, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Tao Li, Yingwen Wu, Sizhe Chen, Kun Fang, and Xiaolin Huang. Subspace adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13409–13418, 2022.

Todd Litman. *Autonomous vehicle implementation predictions*. Victoria Transport Policy Institute Victoria, BC, Canada, 2017.

Guanxiong Liu, Issa Khalil, and Abdallah Khreishah. Using single-step adversarial training to defend iterative adversarial examples. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pp. 17–27, 2021.

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.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Geon Yeong Park and Sang Wan Lee. Reliably fast adversarial training via latent adversarial perturbation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7758–7767, 2021.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pp. 1528–1540, 2016.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pp. 464–472. IEEE, 2017.

Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, et al. Towards efficient and effective adversarial training. *Advances in Neural Information Processing Systems*, 34:11821–11833, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

BS Vivek and R Venkatesh Babu. Regularizers for single-step adversarial training. *arXiv preprint arXiv:2002.00614*, 2020a.

BS Vivek and R Venkatesh Babu. Single-step adversarial training with dropout scheduling. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 947–956. IEEE, 2020b.

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

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

Haizhong Zheng, Ziqi Zhang, Juncheng Gu, Honglak Lee, and Atul Prakash. Efficient adversarial training with transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1181–1190, 2020.

A Experiment with Previous Initialization

Most works build perturbations based on zero or random initialization, but Zheng et al. (2020) and Liu et al. (2021) found that perturbations are highly transferable between models from adjacent epochs, so they used perturbations from adjacent epochs to intensify the effect of perturbations, which is formalized as follows:

$$\eta_t = (\eta_{t-1} + \delta_{t-1}) \cdot \beta, \tag{11}$$

where t is the epoch, η_{t-1} and δ_{t-1} saved from the adjacent epoch and β is the hyperparameter to control the strength of the initialization.

In this part, we will show the effect of our method by using previous initialization by AAER$_{\mathrm{PC}}$: AAER with previous initialization and clipped perturbations and AAER$_{\mathrm{PUC}}$: AAER with previous initialization and unclipped perturbations. We set $\beta = 0.5$ for the previous initialization experiments, and report the results on Cifar10/100 in Table 3 and Table 4.

Table 3: CIFAR10: Accuracy of different methods and different noise magnitudes using PreActResNet-18 under L_∞ threat model. The top number is the natural accuracy while the bottom number is the PGD-50-10 accuracy. The results are averaged over 3 random seeds and reported with the standard deviation.

noise magnitude	8/255	12/255	16/255	32/255
AAER$_{\mathrm{PC}}$ with ATTA	83.68 ± 0.14 47.01 ± 0.29	73.03 ± 1.90 33.56 ± 0.67	62.67 ± 1.32 25.01 ± 0.50	28.39 ± 1.55 13.93 ± 0.60
AAER$_{\mathrm{PC}}$	83.52 ± 0.21 47.14 ± 0.31	74.29 ± 0.63 33.38 ± 0.71	63.52 ± 1.41 24.83 ± 0.75	26.90 ± 1.64 14.21 ± 0.38
AAER$_{\mathrm{PUC}}$ with ATTA	77.25 ± 0.12 51.25 ± 0.10	63.54 ± 0.63 38.39 ± 0.70	45.93 ± 3.66 27.34 ± 1.91	19.28 ± 1.08 9.96 ± 0.96
AAER$_{\mathrm{PUC}}$	77.23 ± 0.55 50.66 ± 0.45	63.68 ± 0.52 38.51 ± 0.29	49.45 ± 1.14 28.85 ± 0.31	23.20 ± 1.21 13.11 ± 0.68

Table 4: CIFAR100: Accuracy of different methods and different noise magnitudes using PreActResNet-18 under L_∞ threat model. The top number is the natural accuracy while the bottom number is the PGD-50-10 accuracy. The results are averaged over 3 random seeds and reported with the standard deviation.

noise magnitude	8/255	12/255	16/255	32/255
AAER$_{\mathrm{PC}}$ with ATTA	57.66 ± 0.28 25.47 ± 0.13	47.44 ± 0.48 16.97 ± 0.04	37.10 ± 0.43 12.00 ± 0.14	20.00 ± 0.22 4.68 ± 0.46
AAER$_{\mathrm{PC}}$	57.66 ± 0.82 25.59 ± 0.41	45.59 ± 0.57 16.53 ± 0.16	35.30 ± 0.23 12.03 ± 0.29	17.41 ± 0.53 5.05 ± 0.03
AAER$_{\mathrm{PUC}}$ with ATTA	50.91 ± 1.08 27.43 ± 1.00	37.93 ± 0.75 18.86 ± 0.48	25.98 ± 1.40 12.81 ± 0.56	9.62 ± 0.51 4.14 ± 0.15
AAER$_{\mathrm{PUC}}$	52.52 ± 0.86 27.94 ± 0.34	36.97 ± 0.87 19.19 ± 0.62	24.90 ± 0.70 13.32 ± 0.36	11.08 ± 0.48 4.43 ± 0.21

From Table 3 and Table 4, we can observe that our method with previous initialization can still successfully achieve high robustness, even achieving better robustness in some settings compared to random initialization. However, using previous initialization has some negative effects on natural accuracy.

β Selection. The hyperparameter β determines the strength of the previous initialization perturbations, and the effect of different β on test accuracy is shown in Figure 5. When β is 0 which is

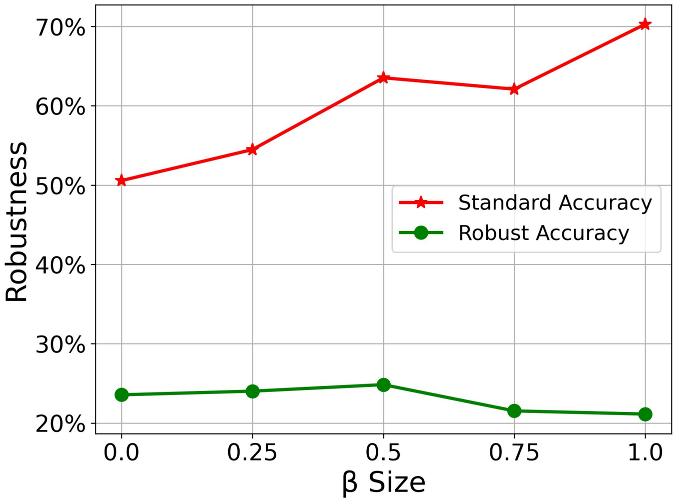

Figure 5: Effect of different size β. The red and green line are the natural and robust test accuracy, respectively. We do this experiment based on AAER$_{\mathrm{PC}}$ with 16/255 noise magnitude.

equivalent to using zero initialization, increasing β leads to higher natural accuracy. When β is greater than 0.5, it is observed increasing β makes model robustness decrease. Therefore, we set β to 0.5 to achieve the best trade-off between natural and robust test accuracy.

Data Argumentation Technique. We notice that Zheng et al. (2020) proposed a data argumentation technique ATTA, which adds different arguments at each epoch. We add this data argumentation technique on AAER$_{\mathrm{PC}}$ as shown in Table 3 and Table 4. We can observe that the ATTA does not or slightly improves our method accuracy, but the training time will significantly increase from 30.5S to 43.1S. Therefore, our method AAER do not use the data argumentation technique ATTA.

B Experiment with WideResNet Architecture

We also compare the performance of our method using WideResNet, which is more complex than PreActResNet. The settings are the same as PreActResNet-18, and we report the results on Cifar10/100 in Table 5 and Table 6.

Table 5: CIFAR10: Accuracy of different methods with 8/255 noise magnitude using WideResNet-34 under L_∞ threat model. The results are averaged over 3 random seeds and reported with the standard deviation.

method	AAER$_{\mathrm{RC}}$	AAER$_{\mathrm{RUC}}$	AAER$_{\mathrm{PC}}$	AAER$_{\mathrm{PUC}}$	PGD-2	PGD-10 (2/255)	PGD-10 (ϵ/10)
natural accuracy	87.83 ± 0.14	82.34 ± 0.23	86.72 ± 0.44	80.82 ± 0.29	88.68 ± 0.14	85.53 ± 0.22	88.51 ± 0.30
robust accuracy	47.54 ± 0.42	52.40 ± 0.53	48.89 ± 0.51	52.89 ± 0.47	47.32 ± 0.50	53.70 ± 0.53	47.72 ± 0.84
training time	227.5S				281.6S	1031.5S	

From Table 5 and Table 6, we can observe that our method can still successfully achieve high robustness in other architectures. Although, the PGD-10 AT seems to better utilize the complex network to achieve higher natural and robust accuracy. However, it is worth noting that complex networks can better reflect the efficiency of our method in terms of training time, while our method can achieve comparable robustness.

Table 6: CIFAR100: Accuracy of different methods with 8/255 noise magnitude using WideResNet-34 under L_∞ threat model. The results are averaged over 3 random seeds and reported with the standard deviation.

method	AAER$_{\text{RC}}$	AAER$_{\text{RUC}}$	AAER$_{\text{PC}}$	AAER$_{\text{PUC}}$	PGD-2	PGD-10 (2/255)	PGD-10 (ϵ/10)
natural accuracy	61.75 ± 0.38	56.73 ± 0.36	61.23 ± 0.24	56.18 ± 0.89	64.64 ± 0.27	60.34 ± 0.34	64.26 ± 0.06
robust accuracy	26.79 ± 0.30	29.89 ± 0.66	27.13 ± 0.10	30.11 ± 0.28	26.47 ± 0.10	30.02 ± 0.09	26.45 ± 0.30
training time	228.5S				285.7S	1036.7S	

C Evaluation Based on Auto Attack

Auto Attack Croce & Hein (2020) is regarded as the most reliable robustness evaluation to date, which is an ensemble of complementary attacks, consisting of three white-box attacks APGD-CE, APGD-DLR, and FAB and a black-box attack Square Attack. We report the results on Cifar10/100 in Table 7 and Table 8.

Table 7: CIFAR10: Accuracy of different methods and different noise magnitudes using PreactResNet-18 under L_∞ threat model. The number is the Auto Attack accuracy while the natural accuracy is same as Table 1. The results are averaged over 3 random seeds and reported with the standard deviation.

noise magnitude	8/255	12/255	16/255	32/255
RS-FGSM	43.17 ± 0.34	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
ATTA	40.09 ± 0.39	16.18 ± 14.03	0.00 ± 0.00	4.90 ± 4.24
FreeAT	40.23 ± 0.33	28.04 ± 0.73	0.00 ± 0.00	0.00 ± 0.00
ZeroGrad	43.48	-	-	-
MulitGrad	44.39	-	-	-
N-FGSM	44.43 ± 0.24	30.32 ± 0.08	19.06 ± 1.81	6.78 ± 0.75
Grad Align	44.82 ± 0.09	30.05 ± 0.17	19.60 ± 0.47	7.89 ± 2.62
AAER$_{\text{RC}}$	43.22 ± 0.24	27.20 ± 0.35	16.91 ± 0.41	10.54 ± 0.67
AAER$_{\text{RUC}}$	46.29 ± 0.23	31.00 ± 0.17	21.16 ± 0.15	10.72 ± 1.74
AAER$_{\text{PC}}$	43.58 ± 0.28	27.22 ± 0.45	16.87 ± 0.38	9.76 ± 0.46
AAER$_{\text{PUC}}$	45.44 ± 0.44	31.01 ± 0.35	21.33 ± 0.22	9.50 ± 0.79
PGD-2	42.97 ± 0.65	28.63 ± 0.38	18.52 ± 0.55	3.77 ± 0.02
PGD-10 (2/255)	46.95 ± 0.54	33.30 ± 0.20	22.29 ± 0.27	2.29 ± 0.10
PGD-10 (ϵ/10)	43.44 ± 0.45	29.82 ± 0.43	19.92 ± 0.63	9.61 ± 0.41

In Table 7 and Table 8, we can observe that our method can still achieve boost adversarial robustness in different adversarial attacks. Surprisingly, the unclipped AAER achieves higher robustness with 32/255 noise magnitude under Auto Attack, which is slightly different from the result under PGD-50-10 attack.

D Experiment with Universal λ

It is worth noting that unlike other SSAT methods (such as Grad Align (Andriushchenko & Flammarion, 2020) and ZeroGrad (Golgooni et al., 2021)), our method can achieve robustness without tuning hyperparameters with different adversarial budgets, the universal λ settings are shown in Ta-

Table 8: CIFAR100: Accuracy of different methods and different noise magnitudes using PreactResNet-18 under L_∞ threat model. The number is the Auto Attack accuracy while the natural accuracy is same as Table 1. The results are averaged over 3 random seeds and reported with the standard deviation.

noise magnitude	8/255	12/255	16/255	32/255
RS-FGSM	7.98 ± 11.91	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
ATTA	20.09 ± 0.06	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
FreeAT	18.28 ± 0.20	12.37 ± 0.14	0.00 ± 0.00	0.00 ± 0.00
ZeroGrad	21.15	-	-	-
MulitGrad	21.62	-	-	-
N-FGSM	22.68 ± 0.25	14.57 ± 0.09	10.30 ± 0.14	0.78 ± 1.36
Grad Align	21.87 ± 0.13	13.78 ± 0.11	9.64 ± 0.12	1.76 ± 0.70
AAER$_{\mathrm{RC}}$	21.79 ± 0.11	12.92 ± 0.20	8.81 ± 0.07	2.66 ± 0.10
AAER$_{\mathrm{RUC}}$	23.77 ± 0.10	14.62 ± 0.24	9.86 ± 0.11	2.94 ± 0.14
AAER$_{\mathrm{PC}}$	21.79 ± 0.64	12.48 ± 0.25	8.38 ± 0.18	2.54 ± 0.13
AAER$_{\mathrm{PUC}}$	23.38 ± 0.27	14.23 ± 0.27	9.16 ± 0.22	2.59 ± 0.30
PGD-2	21.52 ± 0.14	13.69 ± 0.02	9.56 ± 0.07	1.76 ± 0.22
PGD-10 (2/255)	23.78 ± 0.08	15.61 ± 0.09	10.93 ± 0.05	2.18 ± 0.09
PGD-10 (ϵ/10)	21.60 ± 0.03	13.95 ± 0.10	10.18 ± 0.08	3.76 ± 0.10

ble 9 and Table 10. For CIFAR-10, we set $\lambda 1 = 8.0$ $\lambda 2 = 5.0$ for AAER with clipped perturbations, and $\lambda 1 = 6.5$ $\lambda 2 = 5.0$ for AAER with unclipped perturbations. For CIFAR-100, we set $\lambda 1 = 7.5$ $\lambda 2 = 3.0$, and $\lambda 1 = 6.5$ $\lambda 2 = 2.5$ for AAER with clipped and unclipped perturbations, respectively.

Table 9: CIFAR10: The best and universal setting for different noise magnitudes. Last panel is universal λ setting, other panels are best λ setting. The top number is $\lambda 1$ while the bottom number is $\lambda 2$.

Dataset / Method	8/255	12/255	16/255	32/355	Universal
clipped perturbations	6.5 3.0	7.0 4.5	8.0 5.0	9.0 7.5	8.0 5.0
unclipped perturbations	4.0 1.5	4.5 3.0	6.5 5.0	7.0 10.0	6.5 5.0

Table 10: CIFAR100: The best and universal setting for different noise magnitudes. Last panel is universal λ setting, other panels are best λ setting. The top number is $\lambda 1$ while the bottom number is $\lambda 2$.

Dataset / Method	8/255	12/255	16/255	32/355	Universal
clipped perturbations	6.5 2.0	7.0 2.5	7.5 3.0	9.0 1.5	7.5 3.0
unclipped perturbations	4.0 1.0	5.0 2.0	6.5 2.5	7.5 1.5	6.5 2.5

We report the universal λ results on Cifar10/100 in Table 11 and Table 12. We can observe that the results using universal λ can still achieve high robustness on both datasets. The absence of hyperparameter tuning provides our method with unparalleled generality and adaptability.

Table 11: CIFAR10: Accuracy of universal AAER with different noise magnitudes using PreactResNet-18 under L_∞ threat model. The top number is the natural accuracy while the bottom number is the PGD-50-10 accuracy. The results are averaged over 3 random seeds and reported with the standard deviation.

noise magnitude	8/255	12/255	16/255	32/255
AAER$_{\mathrm{RC}}$ ($\lambda 1$ = 8.0, $\lambda 2$ = 5.0)	84.11 ± 0.43 46.04 ± 0.46	74.73 ± 1.19 32.66 ± 0.33	65.88 ± 0.60 24.67 ± 0.52	25.86 ± 0.57 14.64 ± 0.27
AAER$_{\mathrm{RUC}}$ ($\lambda 1$ = 6.5, $\lambda 2$ = 5.0)	76.06 ± 0.59 50.50 ± 0.19	64.26 ± 0.52 38.03 ± 0.12	55.32 ± 0.55 30.31 ± 0.15	24.56 ± 0.71 11.29 ± 1.79
AAER$_{\mathrm{PC}}$ ($\lambda 1$ = 8.0, $\lambda 2$ = 5.0)	83.39 ± 0.20 46.99 ± 0.31	72.15 ± 0.55 33.20 ± 0.47	63.52 ± 1.41 24.83 ± 0.75	27.71 ± 1.65 13.70 ± 0.27
AAER$_{\mathrm{PUC}}$ ($\lambda 1$ = 6.5, $\lambda 2$ = 5.0)	76.03 ± 0.27 50.60 ± 0.08	60.57 ± 1.24 37.01 ± 0.28	49.45 ± 1.14 28.85 ± 0.31	21.21 ± 1.29 12.14 ± 0.72

Table 12: CIFAR100: Accuracy of universal AAER with different noise magnitudes using PreactResNet-18 under L_∞ threat model. The top number is the natural accuracy while the bottom number is the PGD-50-10 accuracy. The results are averaged over 3 random seeds and reported with the standard deviation.

noise magnitude	8/255	12/255	16/255	32/255
AAER$_{\mathrm{RC}}$ ($\lambda 1$ = 7.5, $\lambda 2$ = 3.0)	56.32 ± 0.40 24.86 ± 0.32	47.37 ± 0.63 16.41 ± 0.06	39.44 ± 0.88 12.33 ± 0.14	16.94 ± 0.94 5.12 ± 0.09
AAER$_{\mathrm{RUC}}$ ($\lambda 1$ = 6.5, $\lambda 2$ = 2.5)	49.80 ± 0.47 27.47 ± 0.23	38.56 ± 0.70 18.40 ± 0.68	29.27 ± 0.51 13.96 ± 0.07	10.88 ± 1.22 4.14 ± 0.23
AAER$_{\mathrm{PC}}$ ($\lambda 1$ = 7.5, $\lambda 2$ = 3.0)	55.53 ± 1.05 25.14 ± 0.41	45.20 ± 0.67 16.17 ± 0.44	35.30 ± 0.23 12.03 ± 0.29	13.84 ± 1.05 4.85 ± 0.21
AAER$_{\mathrm{PUC}}$ ($\lambda 1$ = 6.5, $\lambda 2$ = 2.5)	49.04 ± 0.64 27.09 ± 0.17	36.82 ± 1.00 18.53 ± 0.57	24.90 ± 0.70 13.32 ± 0.36	7.16 ± 0.54 3.78 ± 0.09

E Settings and Results on SVHN and Tiny-imagenet

SVHN Settings and Results. For experiments on SVHN, we follow the settings of de Jorge et al. (2022), which use the cyclical learning rate schedule with 15 epochs that reaches its maximum learning rate (0.05) when 40% (6) epochs are passed. We show the best λ settings on SVHN in the Table 13. In Table 14, we show the AAER performance on the SVHN dataset, and the competing baselines result from de Jorge et al. (2022). We can observe that our method can effectively improve the robust accuracy and prevent CO with different noise magnitudes.

Tiny-imagenet Settings and Results. For experiments on Tiny-imagenet, we use the cyclical learning rate schedule with 30 epochs that reaches its maximum learning rate (0.2) when half of the epochs (15) are passed. We set $\lambda 1 = 3.0$ $\lambda 2 = 1.0$ for AAER with clipped perturbations, and $\lambda 1 = 4.0$ $\lambda 2 = 0.5$ for AAER with unclipped perturbations. In Table 15, we show the AAER performance on the Tiny-imagenet dataset. We can observe that our method can effectively improve the robust accuracy and prevent CO.

Table 13: SVHN: The best setting for different noise magnitudes. The top number is $\lambda 1$ while the bottom number is $\lambda 2$.

Dataset / Method	4/255	8/255	12/255
clipped perturbations	3.0 5.0	3.5 6.0	5.0 8.5
unclipped perturbations	1.0 2.5	5.0 4.5	6.5 5.0

Table 14: SVHN: Accuracy of different methods and different noise magnitudes using PreActResNet-18 under L_∞ threat model. The top number is the natural accuracy while the bottom number is the PGD-50-10 accuracy. The results are averaged over 3 random seeds and reported with the standard deviation.

noise magnitude	4/255	8/255	12/255
RS-FGSM	95.09 ± 0.09 71.28 ± 0.4	94.46 ± 0.16 0.0 ± 0.0	92.74 ± 0.5 0.0 ± 0.0
FreeAT	93.66 ± 0.12 71.61 ± 0.75	91.29 ± 4.07 0.01 ± 0.0	92.36 ± 1.0 0.0 ± 0.0
ZeroGrad	94.81 ± 0.16 71.59 ± 0.22	92.42 ± 1.29 35.93 ± 2.73	88.09 ± 0.4 14.14 ± 0.32
MultiGrad	94.71 ± 0.17 71.98 ± 0.26	94.86 ± 0.97 11.49 ± 16.19	94.48 ± 0.19 0.0 ± 0.0
N-FGSM	94.54 ± 0.15 72.53 ± 0.19	89.56 ± 0.49 45.63 ± 0.11	81.48 ± 1.64 26.13 ± 0.81
Grad Align	94.56 ± 0.21 72.12 ± 0.19	90.1 ± 0.34 43.85 ± 0.14	84.01 ± 0.46 23.62 ± 0.41
AAER_{RC}	94.75 ± 0.70 72.00 ± 0.88	91.40 ± 0.85 42.55 ± 0.55	82.18 ± 4.13 22.95 ± 0.51
AAER_{RUC}	93.81 ± 0.26 73.41 ± 0.23	87.11 ± 0.57 46.03 ± 0.25	78.91 ± 0.84 27.44 ± 1.51
PGD-2	94.66 ± 0.1 73.29 ± 0.29	94.63 ± 1.29 20.68 ± 18.56	94.16 ± 0.54 0.02 ± 0.03
PGD-10	94.37 ± 0.13 74.76 ± 0.19	89.67 ± 0.34 53.95 ± 0.55	80.08 ± 0.93 37.65 ± 0.53

Table 15: Tiny-imagenet: Accuracy of different methods with 8/255 noise magnitude using PreActResNet-18 under L_∞ threat model. The results are averaged over 3 random seeds and reported with the standard deviation.

method	AAER_{RC}	AAER_{RUC}	RS-FGSM	N-FGSM	PGD-2
natural accuracy	47.92 ± 0.39	44.38 ± 0.27	52.28 ±2.64	48.16 ± 0.61	46.43 ± 0.35
robust accuracy	19.38 ± 0.14	21.85 ± 0.01	0.00 ± 0.00	20.82 ± 0.40	20.72 ± 0.32