Learnable Boundary Guided Adversarial Training
Supplementary Material

A. Robustness under Black-box attack

Table 7: Comparison of our method with previous defense models under black-box attack on CIFAR-100 and CIFAR-10. To rule out randomness, the numbers are averaged over 2 independently trained models. Acc_n represents accuracy on natural images. $BAcc_r$ represents robustness under black-box attack. $WAcc_r$ represents robustness under white-box attack

Target Models	**BAcc$_r$**	**WAcc$_r$**	**Acc$_n$**	**Source Models**	**Dataset**
TRADES ($\alpha = 1$)	61.29%	25.31%	62.37%	Natural	CIFAR-100
TRADES ($\alpha = 6$)	55.52%	30.93%	56.51%	Natural	CIFAR-100
LBGAT+ALP	61.38%	**35.25%**	62.67%	Natural	CIFAR100
LBGAT+TRADES (α=0)	**68.35%**	**33.01%**	**70.03%**	Natural	CIFAR-100
TRADES ($\alpha = 1$)	42.32%	25.31%	62.37%	LBGAT+TRADES (α=0)	CIFAR-100
TRADES ($\alpha = 6$)	41.67%	30.93%	56.51%	LBGAT+TRADES (α=0)	CIFAR-100
LBGAT+ALP	45.68%	**35.25%**	62.67%	TRADES ($\alpha = 6$)	CIFAR-100
LBGAT+TRADES (α=0)	**50.27%**	**33.01%**	**70.03%**	TRADES ($\alpha = 6$)	CIFAR-100
TRADES ($\alpha = 1$)	87.00%	49.14%	**88.64%**	Natural	CIFAR-10
TRADES ($\alpha = 6$)	83.30%	56.61%	84.92%	Natural	CIFAR-10
LBGAT+TRADES ($\alpha = 0$)	**87.20%**	**57.55%**	**88.22%**	Natural	CIFAR-10
TRADES ($\alpha = 1$)	66.18%	49.14%	**88.64%**	LBGAT+TRADES(α=0)	CIFAR-10
TRADES ($\alpha = 6$)	67.18%	56.61%	84.92%	LBGAT+TRADES (α=0)	CIFAR-10
LBGAT+TRADES ($\alpha = 0$)	**68.45%**	**57.55%**	**88.22%**	TRADES (α=6)	CIFAR-10

B. Our Method Creates New SOTA Under the Strongest Auto-Attack on CIFAR-100

To further show the effectiveness of our method, we compare with more previous works. The experimental results are shown in Table 8. On the more challlenging CIFAR-100 dataset, our method creates a new state-of-the-art (SOTA) on both robustness and natural accuracy. Specifically, our LBGAT ($\alpha = 0$) model with WideResNet-34-10 architecture significantly outperforms previous SOAT method [6] by 7.08% in the aspect of performance on natural data. Meanwhile, our method surpasses it with respect to model robustness. Further, our strongest model LBGAT ($\alpha = 6$) with WideResNet-34-10 architecture enjoys 2.4% higher robustness than [6].

Moreover, It is worthy to note that our LBGAT ($\alpha = 6$) model achieves even strong robustness than the model, by Hendrycks *et al*. [18], pre-trained on full ImageNet. At the same time, we also surpasses it in the aspect of natural accuracy.

Table 8: More comparisons under the strongest Auto-Attack on CIFAR-100 dataset. "†" denotes numbers are directly copied from [10]. "⋆" denotes that the method has used additional unlabeled data.

Methods	Model	**Acc$_n$**	**Acc$_r$**
LBGAT ($\alpha = 0$) Ours	WideResNet-34-20	**71.00%**	**27.66%**
LBGAT ($\alpha = 6$) Ours	WideResNet-34-20	62.55%	**30.20%**
LBGAT ($\alpha = 0$) Ours	WideResNet-34-10	**70.03%**	**27.05%**
LBGAT ($\alpha = 6$) Ours	WideResNet-34-10	60.43%	29.34%
TRADES ($\alpha = 1$) [56]	WideResNet-34-10	62.37%	22.24%
TRADES ($\alpha = 6$) [56]	WideResNet-34-10	56.50%	26.87%
Sitawarin *et al*. [38] †	WideResNet-34-10	62.82%	24.57%
Chen *et al*. [6] †	WideResNet-34-10	62.15%	26.94%
Hendrycks *et al*. [18] †⋆	WideResNet-28-10	59.23%	28.42%
Rice *et al*. [33] †	ResNet-18	53.83%	18.95%

C. More Comparisons Under the Strongest Auto-Attack on CIFAR-10

We also compare with more previous methods on CIFAR-10 dataset. The experimental results are summarized in Table 9. Our LBGAT ($\alpha = 0$) model with WideResNet-34-10 architecture can consistently enjoy higher natural performance while keeping the strongest robustness. We observe that though many fast adversarial training methods, like [45, 35] are proposed to accelerate the training process, their performance are usually unsatisfied.

Table 9: More comparisons under the strongest Auto-Attack on CIFAR-10 dataset. "†" denotes numbers are directly copied from [10]. "∗" denotes the methods aiming to accelerate adversarial training.

Methods	Model	**Acc$_n$**	**Acc$_r$**
LBGAT ($\alpha = 0$) Ours	WideResNet-34-20	**88.70%**	**53.58%**
LBGAT ($\alpha = 6$) Ours	WideResNet-34-20	83.61%	**54.45%**
LBGAT ($\alpha = 0$) Ours	WideResNet-34-10	**88.22%**	**52.86%**
LBGAT ($\alpha = 6$) Ours	WideResNet-34-10	81.98%	53.14%
Rice *et al*. [33] †	WideResNet-34-20	85.34%	53.42%
TRADES ($\alpha = 1$)	WideResNet-34-10	**88.64%**	48.11%
TRADES ($\alpha = 6$)	WideResNet-34-10	84.92%	52.64%
Kumari *et al*. [22] †	WideResNet-34-10	87.80%	49.12%
Mao *et al*. [28] †	WideResNet-34-10	86.21%	47.41%
Zhang *et al*. [55] †∗	WideResNet-34-10	87.20%	44.83%
Shafahi *et al*. [35] †∗	WideResNet-34-10	86.11%	41.47%
Chan *et al*. [5] †	WideResNet-34-10	**93.79%**	0.26%
Wang *et al*. [45] †∗	WideResNet-28-10	**92.80%**	29.35%
Qin *et al*. [32] †	WideResNet-40-8	86.28%	52.81%
Chen *et al*. [8] †	ResNet-50	86.04%	51.56%
Xiao *et al*. [47] †	DenseNet-121	79.28%	18.50%
Wong *et al*. [46] †	ResNet-18	83.34%	43.21%