A Appendix

A.1 Restricted Imagenet Result

In addition to CIFAR, we also test the performance on the Restricted Imagenet dataset, which has been used in previous papers such as (Sinha et al., 2019). This dataset consists of a subset of imagenet classes which have been grouped into 9 different classes. The experimental results are shown in Table 5. All the results except our methods are reported in (Sinha et al., 2019).

Table 7: The clean and robust accuracy of Resnet50 models trained by various defense methods. All robust accuracy are measured under $\epsilon = 8/255$ ℓ_∞ ball. We reported the best performance listed in the papers. $^{(*)}$ denotes random-restart is applied in the testing attack. $^{(X)}$ denotes it use a X-step PGD attack

Methods	Clean accuracy	PGD accuracy	C&W accuracy
Adversarial training [20]	91.83%	17.52%	✗
FAT (Sinha et al., 2019)	**91.59 %**	18.81%	✗
LAT (Sinha et al., 2019)	89.86 %	22.00%	✗
CAT (ours)	88.63%	58.4%$^{(*20)}$	58.4%$^{(*20)}$

A.2 Omitted Proofs

In this section we provide the omitted proof for Theorem 3.2, which is adapted from Theorem 2.1 from [33]. They defined the all layer margin for a k-layer network $h_\theta(\boldsymbol{x}) = f_k \circ f_{k-1} \circ \cdots f_1(\boldsymbol{x})$ and perturbation $\delta = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \cdots \boldsymbol{\delta}_k)$ as follows:

$$\begin{aligned} h_1(\boldsymbol{x}, \delta) =& f_1(\boldsymbol{x}) + \boldsymbol{\delta}_1 \|\boldsymbol{x}\|_2 \\ h_i(\boldsymbol{x}, \delta) =& f_i(h_{i-1}(\boldsymbol{x}, \boldsymbol{\delta})) + \boldsymbol{\delta}_i \|h_{i-1}(\boldsymbol{x}, \delta)\|_2 \\ H_\theta(\boldsymbol{x}, \delta) =& h_k(\boldsymbol{x}, \delta). \end{aligned}$$

They define the all-layer margin as the minimum norm of $\delta = (\boldsymbol{\delta}_i)_{i=1}^k$ required that causes the classifier to make a false prediction.

$$\begin{aligned} m_F(\boldsymbol{x}, y) := \min_{\boldsymbol{\delta}^i, \boldsymbol{\delta}^o} & \sqrt{\|\boldsymbol{\delta}^i\|^2 + \|\boldsymbol{\delta}^o\|^2} \\ & \text{subject to } \max_{y'} H_\theta(\boldsymbol{x}, \boldsymbol{\delta}^i, \boldsymbol{\delta}^o)_{y'} \neq y. \end{aligned} \tag{8}$$

They consider the function class $\mathcal{F} = \{f_k \circ f_{k-1} \cdots \circ f_1 : f_i \in \mathcal{F}_i\}$ be the class of compositions of functions from function classes $\mathcal{F}_1, \cdots \mathcal{F}_k$. They achieve the generalization bound as follows:

Theorem A.1 (Theorem 2.1 from [33]) *In the above setting, with probability $1 - \delta$ over the draw of the data, all classifiers $F \in \mathcal{F}$ which achieve training error 0 satisfy*

$$\mathbb{E}[f_\theta(\boldsymbol{x}) = y] \lesssim \frac{\sum_i \mathcal{C}_i \log^2 n}{\sqrt{n}} \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{1}{m_F(\boldsymbol{x}_i, y_i)}} + \zeta,$$

where ζ is of small order $O\left(\frac{1}{n} \log(1/\delta)\right)$.

For our problem, we define $h_\theta(\boldsymbol{x}) := f_2 \circ f_1(\boldsymbol{x})$, where f_1 is identity mapping, and f_2 is the original function h_θ. Therefore the all layer margin is reduced to our bilateral margin:

$$\begin{aligned} h_1(\boldsymbol{x}, \boldsymbol{\delta}^i) =& f_1(\boldsymbol{x}) + \boldsymbol{\delta}^i \|\boldsymbol{x}\|_2 = \boldsymbol{x} + \boldsymbol{\delta}^i \|\boldsymbol{x}\| \\ H_\theta(\boldsymbol{x}, \boldsymbol{\delta}) = h_2(\boldsymbol{x}, \boldsymbol{\delta}^i, \boldsymbol{\delta}^o) =& f_2(h_1(\boldsymbol{x}, \boldsymbol{\delta}^i)) + \boldsymbol{\delta}^o \|h_1(\boldsymbol{x}, \boldsymbol{\delta}^i)\| \\ =& h_\theta(\boldsymbol{x} + \boldsymbol{\delta}^i \|\boldsymbol{x}\|) + \boldsymbol{\delta}^o \|\boldsymbol{x} + \boldsymbol{\delta}^i \|\boldsymbol{x}\|\|. \end{aligned}$$

Next, notice since f_1 is identity mapping, and composition with h_θ doesn't affect the overall complexity. We apply Theorem A.1 and get our result.

A.3 Performance under different PGD strength

For CIFAR10, we evaluate all the models under different tolerance of white-box ℓ_∞-norm bounded non-targeted PGD and C&W attack.

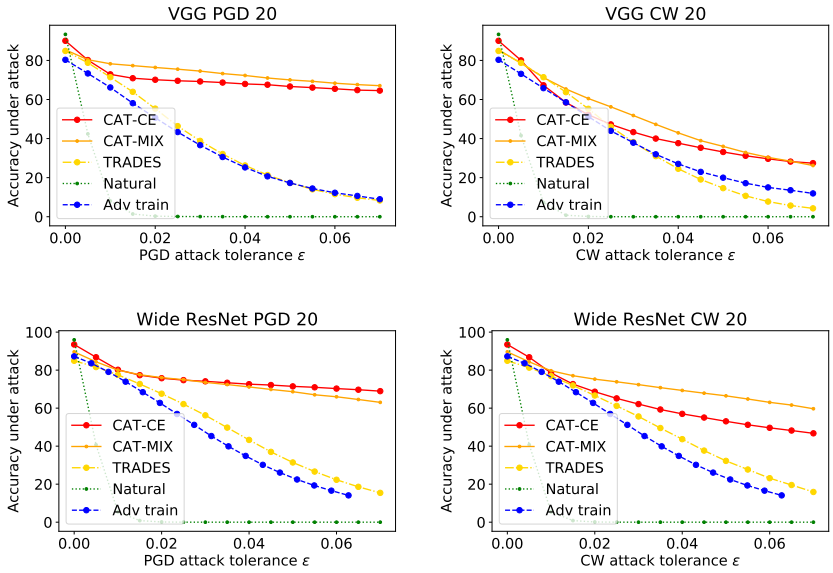
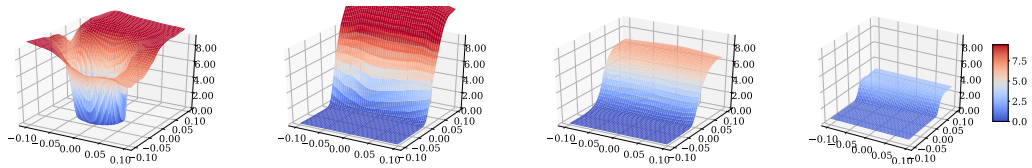


Figure 2: Robust accuracy under different levels of attacks on CIFAR-10 datasets with VGG and Wide-ResNet architectures. CAT-CE and CAT-MIX clearly outperform TRADES and adversarial training.

A.4 Loss Landscape Exploration



(a) Natural (b) Adv train (c) TRADES (d) CAT

Figure 3: Loss landscape comparison of different adversarial training methods

To further verify the superior robustness using CAT, we visualize the loss landscape of different training methods in Figure 3. Following the implementation in [11], we divide the data input along a linear space grid defined by the sign of the input gradient and a random Rademacher vector, where the x- and y- axes represent the magnitude of the perturbation added in each direction and the z-axis represents the loss. As shown in Figure 3, CAT generates a model with a lower and smoother loss landscape. Also, it could be taken as another strong evidence that we have found a robust model through CAT training.