

Improving the Accuracy-Robustness Trade-Off of Classifiers via Adaptive Smoothing*

Yatong Bai[†], Brendon G. Anderson[†], Aerin Kim[‡], and Somayeh Sojoudi[†]

Abstract. While prior research has proposed a plethora of methods that build neural classifiers robust against adversarial robustness, practitioners are still reluctant to adopt them due to their unacceptably severe clean accuracy penalties. Real-world services based on neural networks are thus still unsafe. This paper significantly alleviates the accuracy-robustness trade-off by mixing the output probabilities of a standard classifier and a robust classifier, where the standard network is optimized for clean accuracy and is not robust in general. We show that the robust base classifier’s confidence difference for correct and incorrect examples is the key to this improvement. In addition to providing empirical evidence, we theoretically certify the robustness of the mixed classifier under realistic assumptions. We then adapt an adversarial input detector into a mixing network that adaptively adjusts the mixture of the two base models, further reducing the accuracy penalty of achieving robustness. The proposed flexible mixture-of-experts framework, termed “adaptive smoothing,” works in conjunction with existing or even future methods that improve clean accuracy, robustness, or adversary detection. We use strong attack methods, including AutoAttack and adaptive attacks, to evaluate our models’ robustness. On the CIFAR-100 dataset, we achieve an 85.21% clean accuracy while maintaining a 38.72% ℓ_∞ -AutoAttacked ($\epsilon = 8/255$) accuracy, becoming the second most robust method on the RobustBench benchmark as of submission, while improving the clean accuracy by 10 percentage points over all listed models. Code implementation is available at <https://github.com/Bai-YT/AdaptiveSmoothing>.

Key words. machine learning, deep learning, neural networks, adversarial robustness, computer vision

MSC codes. 68T07, 68T05, 68T45, 90C17

DOI. 10.1137/23M1564560

1. Introduction. Neural networks are vulnerable to adversarial attacks in various applications, including computer vision and audio [63, 37], natural language processing [36], and control systems [45]. Due to the widespread application of neural classifiers, ensuring their reliability in practice is paramount.

To mitigate this susceptibility, researchers have explored “adversarial training” (AT) and its improved variants [54, 37, 16, 17, 92], building empirically robust models by training with adversarial examples. Meanwhile, theoretical research has also considered certifying

*Received by the editors April 28, 2023; accepted for publication (in revised form) March 7, 2024; published electronically August 14, 2024. This work is an extension of 6th Annual Learning for Dynamics & Control Conference (L4DC 2024).

<https://doi.org/10.1137/23M1564560>

Funding: This work was supported by grants from ONR, NSF, UC Noyce Initiative, and C3 AI Digital Transformation Institute.

[†]Department of Mechanical Engineering and Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94709 USA (yatong_bai@berkeley.edu, bganderson@berkeley.edu, sojoudi@berkeley.edu).

[‡]Scale AI, San Francisco, CA 94103 USA (aerinkim@gmail.com).

(i.e., mathematically guaranteeing) the robustness of neural classifiers against adversarial perturbations within a radius [7, 60, 9]. “Randomized smoothing” (RS) is one such method that achieves certified robustness with an already trained model at inference time [28, 56]. Improved variants of RS incorporate dimension reduction methods [71] and denoising modules [20]. Recent work [8] has demonstrated that a data-driven locally biased smoothing approach can improve over traditional data-blind RS. However, this method is limited to the binary classification setting and suffers from the performance bottleneck of its underlying one-nearest-neighbor classifier.

Despite the emergence of these proposed remedies to the adversarial robustness issue, many practitioners are reluctant to adopt them. As a result, existing publicly available services are still vulnerable [46, 19], presenting severe safety risks. One important reason for this reluctance is the potential for significantly reduced model performance on clean data. Specifically, some previous works have suggested a fundamental trade-off between accuracy and robustness [83, 90]. Since the sacrifice in unattacked performance is understandably unacceptable in real-world scenarios, developing robust classifiers with minimal clean accuracy degradation is crucial.

Fortunately, recent research has argued that it should be possible to simultaneously achieve robustness and accuracy on benchmark datasets [87]. To this end, variants of adversarial training that improve the accuracy-robustness trade-off have been proposed, including TRADES [90], Interpolated Adversarial Training [55], Instance Adaptive Adversarial Training (IAAT) [18], and many others [26, 13, 72, 85, 89, 82]. However, despite these improvements, degraded clean accuracy is often an inevitable price of achieving robustness. Moreover, standard nonrobust models often achieve enormous performance gains by pretraining on larger datasets with self- or semisupervision [41, 15]. In contrast, the effect of pretraining on robust classifiers is less understood and may be less prominent [24, 34]. As a result, the performance gap between these existing works and the possibility guaranteed in [87] is still huge.

This work builds upon locally biased smoothing [8] and makes a theoretically disciplined step towards reconciling adversarial robustness and clean accuracy, significantly closing this performance gap and thereby providing practitioners additional incentives for deploying robust models. This paper is organized as follows.

- In section 3, observing that the K -nearest-neighbor (K -NN) classifier, a crucial component of locally biased smoothing, becomes a performance bottleneck, we replace it with a robust neural network that can be obtained via various existing methods, and we propose a new smoothing formulation accordingly. The resulting formulation (3.4) is a convex combination of the output probabilities of a standard neural network and a robust one. When the robust neural network has a certified Lipschitz constant or is based on randomized smoothing, the mixed classifier also has a certified robust radius. These contents are presented in our conference paper [14], but are strengthened in this paper.
- In section 4, we propose adaptive smoothing, which adaptively adjusts the mixture of a standard model and a robust model by adopting a type of adversary detector as a “mixing network.” The mixing network controls the convex combination of the output probabilities from the two base networks, further improving the accuracy-robustness trade-off, making the resulting model a mixture-of-experts design. We empirically verify the robustness of the proposed method using gray-box and white-box projected

gradient descent (PGD) attack, AutoAttack, and adaptive attacks, demonstrating that the mixing network is robust against the attack types it is trained with. When the mixing network is trained with a carefully designed adaptive AutoAttack, the composite model significantly gains clean accuracy while sacrificing little robustness. This section and the corresponding experiment results are entirely new relative to our conference paper [14], and are crucial for achieving the much improved accuracy-robustness trade-off over existing works.

Compared to existing methods for improving the accuracy-robustness trade-off, most of which are training-based, adaptive smoothing has several key advantages:

- Adaptive smoothing is agnostic to how the standard and robust base models are trained. Hence, one can quickly swap the base classifiers with already trained standard or robust models. Therefore, our method is highly versatile and can be coupled with existing training-based trade-off improving methods.
- Adaptive smoothing can thus take advantage of pretraining on large datasets via the standard base classifier and benefit from ongoing advancements in robust training methods via the robust base model. Meanwhile, training-based methods have limited compatibilities, since they may conflict with certain techniques essential to achieving state-of-the-art (SOTA) clean or robust accuracy. As a result, adaptive smoothing achieves better results: it significantly boosts clean accuracy while maintaining near-SOTA robustness.
- Adaptive smoothing allows for an interpretable continuous adjustment between accuracy and robustness at inference time, which can be achieved by simply adjusting the mixture ratio. On the other hand, not all training-based methods allow for this adjustment. For those that do, this adjustment involves training an entirely new robust model.
- When the mixing ratio is fixed and the robust base model has a certified robust radius with a nonzero margin, the mixed classifier can be certified. Since certified models are often also certifiable with a nonzero margin, this condition is commonly satisfied in practice. For empirically robust base classifiers that are not certifiable, an estimation can be performed.

During the reviewing period of this paper, the authors of [57] verified that our mixed classifier simultaneously improves the clean accuracy and the robustness against out-of-distribution (OOD) adversarial attacks (i.e., the threat model differs between training and evaluation), achieving state-of-the-art OOD adversarial robustness among a plethora of models, including the robust base classifier of our mixed classifier. This observation further bolsters the thesis that our proposed method achieves the accuracy-robustness trade-off.

2. Background and related works.

2.1. Notation. The symbol $\|\cdot\|_p$ denotes the ℓ_p norm of a vector and $\|\cdot\|_{p*}$ denotes its dual norm. For a scalar a , $\text{sgn}(a) \in \{-1, 0, 1\}$ denotes its sign. For a natural number c , $[c]$ represents $\{1, 2, \dots, c\}$. For an event A , the indicator function $\mathbb{I}(A)$ evaluates to 1 if A takes place and to 0 otherwise. The probability for an event $A(X)$ to occur is denoted by $\mathbb{P}_{X \sim \mathcal{S}}[A(X)]$, where X is a random variable drawn from the distribution \mathcal{S} .

Consider a model $g : \mathbb{R}^d \rightarrow \mathbb{R}^c$, whose components are $g_i : \mathbb{R}^d \rightarrow \mathbb{R}$, $i \in [c]$, where d is the dimension of the input and c is the number of classes. A classifier $f : \mathbb{R}^d \rightarrow [c]$ can be obtained

via $f(x) \in \arg \max_{i \in [c]} g_i(x)$. In this paper, we assume that $g(\cdot)$ does not have the desired level of robustness, and we refer to it as a “standard classifier” (as opposed to a “robust classifier” which we denote as $h(\cdot)$). Throughout this paper, we regard $g(\cdot)$ and $h(\cdot)$ as the base classifier logits. To denote their output probabilities, we use $\sigma \circ g(\cdot)$ and $\sigma \circ h(\cdot)$. Similarly, $\sigma \circ g_i(\cdot)$ denotes the predicted probability of the i th class from $g(\cdot)$. Moreover, we use \mathcal{D} to denote the set of all validation input-label pairs (x_i, y_i) .

We consider ℓ_p -norm-bounded attacks on differentiable neural networks. A classifier $f(\cdot)$ is considered robust against adversarial perturbations at some input data $x \in \mathbb{R}^d$ if it assigns the same label to all perturbed inputs $x + \delta$ such that $\|\delta\|_p \leq \epsilon$, where $\epsilon \geq 0$ is the attack radius. We use PGD_T to denote the T -step PGD attack.

2.2. Related adversarial attacks and defenses. The fast gradient sign method (FGSM) and PGD attacks based on the first-order maximization of the cross-entropy loss have traditionally been considered classic and straightforward attacks [61, 37]. However, these attacks have been shown to be insufficient as defenses designed against them are often easily circumvented [22, 12, 11, 69]. To this end, various attack methods based on alternative loss functions, Expectation Over Transformation, and black-box perturbations have been proposed. Such efforts include MultiTargeted attack loss [40], AutoAttack [31], adaptive attack [81], minimal distortion attack [32], and many others, even considering attacking test-time defenses [30]. The diversity of attack methods has led to the creation of benchmarks such as RobustBench [29] and ARES-Bench [58] to unify the evaluation of robust models.

On the defense side, while adversarial training [61] and TRADES [90] have seen enormous success, such methods are often limited by a significantly larger amount of required training data [76]. Initiatives that construct more effective training data via data augmentation [73, 38, 39] and generative models [77, 86] have successfully produced more accurate and robust models. Improved versions of adversarial training [47, 84, 78, 66] have also been proposed.

Previous research has developed models that improve robustness by dynamically changing at test time. Specifically, Input-Adaptive Inference improves the accuracy-robustness trade-off by appending side branches to a single network, allowing for early-exit predictions [44]. Other initiatives that aim to enhance the accuracy-robustness trade-off include using the SCORE attack during training [67] and applying adversarial training for regularization [93].

Moreover, ensemble-based defenses, such as random ensemble [59], diverse ensemble [68, 3, 1], and Jacobian ensemble [27], have been proposed. In comparison, this work is distinct in that our mixing scheme uses two separate classifiers, incorporating one nonrobust component while still ensuring the adversarial robustness of the overall design. By doing so, we take advantage of the high performance of modern pretrained models, significantly alleviating the accuracy-robustness trade-off and achieving much higher overall performances. Additionally, unlike some previous ensemble initiatives, our formulation is deterministic and straightforward (in the sense of gradient propagation), making it easier to evaluate its robustness properly. The work [53] also explored assembling an accurate classifier and a robust classifier, but the method considered robustness against distribution shift in a nonadversarial setting and was based on different intuitions. After the submission of this paper, the work [91] also considered leveraging the power of a pair of standard and robust classifiers. However, instead of mixing

the outputs, the authors proposed to distill a new model from the two base classifiers. While this approach also yielded impressive results, the distillation process is time-consuming.

2.3. Locally biased smoothing. Randomized smoothing, popularized by [28], achieves robustness at inference time by replacing the standard classifier $f(\cdot)$ with the smoothed model

$$\tilde{f}(x) \in \arg \max_{i \in [c]} \mathbb{P}_{\delta \sim \mathcal{S}}[f(x + \delta) = i],$$

where \mathcal{S} is a smoothing distribution, for which a common choice is a Gaussian distribution.

Note that \mathcal{S} is independent of the input x and is often zero-mean. The authors of [8] have shown that data-invariant smoothing enlarges the region of the input space at which the prediction of $\tilde{f}(\cdot)$ stays constant. Such an operation may unexpectedly degrade both clean and robust accuracy (the limiting case is when $\tilde{f}(\cdot)$ becomes a constant classifier). Furthermore, when $f(\cdot)$ is a linear classifier, the zero-mean restriction on \mathcal{S} leaves $f(\cdot)$ unchanged. That is, randomized smoothing with a zero-mean distribution cannot help robustify even the most simple linear classifiers. To overcome these limitations, [8] allowed \mathcal{S} to be input-dependent (denoted by \mathcal{S}_x) and nonzero-mean and searched for distributions \mathcal{S}_x that best robustify $\tilde{f}(\cdot)$ with respect to the data distribution. The resulting scheme is “locally biased smoothing.”

It is shown in [8] that, up to a first-order linearization of the base classifier, the optimal locally biased smoothing distribution \mathcal{S}_x shifts the input point in the direction of its true class. Formally, for a binary classifier of the form $f(x) = \text{sign}(g(x))$ with continuously differentiable $g(\cdot)$, maximizing the robustness of $\tilde{f}(\cdot)$ around x over all distributions \mathcal{S}_x with bounded mean yields the optimal locally biased smoothing classifier given by

$$\tilde{f}(x) = \text{sign}(\tilde{g}(x)), \quad \text{where} \quad \tilde{g}(x) = g(x) + \gamma y(x) \|\nabla g(x)\|_{p^*},$$

where $y(x) \in \{-1, 1\}$ is the true class of x , and where $\gamma \geq 0$ is the (fixed) bound on the distribution mean (i.e., $\|\mathbb{E}_{\delta \sim \mathcal{S}_x}[\delta]\|_p \leq \gamma$).

Intuitively, this optimal locally biased smoothing classifier shifts the input along the direction $\nabla g(x)$ when $y(x) = 1$ as a means to make the classifier more likely to label x into class 1, and conversely shifts the input along the direction $-\nabla g(x)$ when $y(x) = -1$. Of course, during inference, the true class $y(x)$ is generally unavailable, and therefore [8] uses a “direction oracle” $h(x) \in \{-1, 1\}$ as a surrogate for $y(x)$, resulting in the locally biased smoothing classifier

$$(2.1) \quad f^\gamma(x) = \text{sign}(h^\gamma(x)), \quad \text{where} \quad h^\gamma(x) = g(x) + \gamma h(x) \|\nabla g(x)\|_{p^*}.$$

Notice that unlike randomized smoothing, the computation (2.1) is deterministic, which is a consequence of the closed-form optimization over \mathcal{S}_x .

In contrast to data-invariant randomized smoothing, the direction oracle $h(\cdot)$ is learned from data, incorporating the data distribution into the manipulation of the smoothed classifier’s decision boundaries. This allows for increases in nonlinearity when the data implies that such nonlinearities are beneficial for robustness, resolving a fundamental limitation of data-invariant smoothing. In general, the direction oracle should come from an inherently robust classifier. Since such a robust model $h(\cdot)$ is often less accurate, the value γ can be viewed as a

trade-off parameter, as it encodes the amount of trust into the direction oracle. The authors of [8] showed that when the direction oracle is a one-nearest-neighbor classifier, locally biased smoothing outperforms traditional randomized smoothing in binary classification.

2.4. Adversarial input detectors. Adversarial inputs can be detected via various methods. For example, [62] proposed to append an additional detection branch to an existing neural network and use adversarial data to train the detector in a supervised fashion. However, Carlini and Wagner [21] showed that it is possible to bypass this detection method. They constructed adversarial examples via the C&W attacks [22] and simultaneously targeted the classification branch and the detection branch by treating the two branches as an “augmented classifier.” According to [21], the detector is effective against the types of attack that it is trained with, but not necessarily the attack types that are absent in the training data. It is thus reasonable to expect the detector to be able to detect a wide range of attacks if it is trained using sufficiently diverse types of attacks (including those targeting the detector itself). While exhaustively covering the entire adversarial input space is intractable, and it is unclear to what degree one needs to diversify the attack types in practice, our experiments show that our modified architecture based on [62] can recognize the SOTA AutoAttack adversaries with a high success rate.

The literature has also considered alternative detection methods that mitigate the above challenges faced by detectors trained in a supervised fashion [23]. Such initiatives include unsupervised detectors [4, 5] and reattacking [2]. Since universally effective detectors have not yet been discovered, the current paper focuses on transferring the properties of the existing detector toward better overall robustness. Future advancements in the field of adversary detection can further enhance the performance of our method.

3. Using a robust neural network as the smoothing oracle. Locally biased smoothing was designed for binary classification, restricting its practicality. Here, we first extend it to the multiclass setting by treating the output $h_i^\gamma(x)$ of each class independently, giving rise to

$$(3.1) \quad h_{\text{smo1},i}^\gamma(x) := g_i(x) + \gamma h_i(x) \|\nabla g_i(x)\|_{p^*} \quad \forall i \in [c].$$

Note that if $\|\nabla g_i(x)\|_{p^*}$ is large for some i , then $h_{\text{smo1},i}^\gamma(x)$ can be large even if both $g_i(x)$ and $h_i(x)$ are small, potentially leading to incorrect predictions. To remove the effect of the magnitude difference across the classes, we propose a normalized formulation as follows:

$$(3.2) \quad h_{\text{smo2},i}^\gamma(x) := \frac{g_i(x) + \gamma h_i(x) \|\nabla g_i(x)\|_{p^*}}{1 + \gamma \|\nabla g_i(x)\|_{p^*}} \quad \forall i \in [c].$$

The parameter γ adjusts between clean accuracy and robustness. It holds that $h_{\text{smo2},i}^\gamma(x) \equiv g_i(x)$ when $\gamma = 0$, and $h_{\text{smo2},i}^\gamma(x) \rightarrow h_i(x)$ when $\gamma \rightarrow \infty$ for all x and all i .

With the mixing procedure generalized to the multiclass setting, we now discuss the choice of the smoothing oracle $h_i(\cdot)$. While K -NN classifiers are relatively robust and can be used as the oracle, their representation power is too weak. On the CIFAR-10 image classification task [52], K -NN only achieves around 35% accuracy on clean test data. In contrast, an adversarially trained ResNet [42] can reach 50% accuracy on attacked test data [61]. This lackluster performance of K -NN becomes a significant bottleneck in the accuracy-robustness

trade-off of the mixed classifier. To this end, we replace the K -NN model with a robust neural network. The robustness of this network can be achieved via various methods, including adversarial training, TRADES, and traditional randomized smoothing.

Further scrutinizing (3.2) leads to the question of whether $\|\nabla g_i(x)\|_{p^*}$ is the best choice for adjusting the mixture of $g(\cdot)$ and $h(\cdot)$. This gradient magnitude term is a result of the setting of $h(x) \in \{-1, 1\}$ considered in [8]. Here, we assume a different setting, where both $g(\cdot)$ and $h(\cdot)$ are multiclass and differentiable. Thus, we further generalize the formulation to

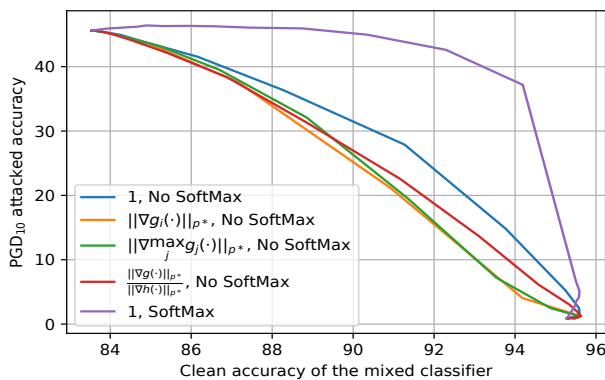
$$(3.3) \quad h_{\text{smo},3,i}^\gamma(x) := \frac{g_i(x) + \gamma R_i(x) h_i(x)}{1 + \gamma R_i(x)} \quad \forall i \in [c],$$

where $R_i(x)$ is an extra scalar term that can potentially depend on both $\nabla g_i(x)$ and $\nabla h_i(x)$ to determine the “trustworthiness” of the base classifiers. Here, we empirically compare four options for $R_i(x)$, namely, 1, $\|\nabla g_i(x)\|_{p^*}$, $\|\nabla \max_j g_j(x)\|_{p^*}$, and $\frac{\|\nabla g_i(x)\|_{p^*}}{\|\nabla h_i(x)\|_{p^*}}$. In subsection SM2.1 in the accompanying supplementary material, we explain how these four options were designed.

Another design choice is whether $g(\cdot)$ and $h(\cdot)$ should be the pre-softmax logits or the post-softmax probabilities. Note that since most attack methods are designed based on logits, the output of the mixed classifier should be logits rather than probabilities. This is because feeding output probabilities into attacks designed around logits effectively results in a redundant softmax layer, which can cause gradient masking, an undesirable phenomenon that makes it hard to evaluate the proposed method’s robustness properly. Thus, we have the following two options that make the mixed model compatible with existing gradient-based attacks:

1. Use the logits for both base classifiers, $g(\cdot)$ and $h(\cdot)$.
2. Use the probabilities for both base classifiers, and then convert the mixed probabilities back to logits. The required “inverse-softmax” operator is simply the natural logarithm.

Figure 1 visualizes the accuracy-robustness trade-off achieved by mixing logits or probabilities with different $R_i(x)$ options. Here, the base classifiers are a pair of standard and adversarially trained ResNet-18s. This “clean accuracy versus PGD₁₀-attacked accuracy” plot concludes that $R_i(x) = 1$ optimizes the accuracy-robustness trade-off, and $g(\cdot)$ and $h(\cdot)$



- “No Softmax” represents Option 1, i.e., use the logits $g(\cdot)$ and $h(\cdot)$.
- “Softmax” represents Option 2, i.e., use the probabilities $\sigma \circ g(\cdot)$ and $\sigma \circ h(\cdot)$.
- With the best formulation, high clean accuracy can be achieved with very little sacrifice on robustness.

Figure 1. Compare the “attacked accuracy–clean accuracy” curves for various $R_i(x)$ options.

should be probabilities. Subsection SM2.2 confirms this selection by repeating Figure 1 with different model architectures, other robust base model training methods, and various attack budgets.

Our selection of $R_i(x) = 1$ differs from $R_i(x) = \|g_i(x)\|_{p*}$ used in [8]. Intuitively, [8] used linear classifier examples to motivate estimating the trustworthiness of the base models with their gradient magnitudes. However, when the base classifiers are highly nonlinear neural networks as in our case, while the local Lipschitzness of a base classifier still correlates with its robustness, its gradient magnitude is not always a good estimator of the local Lipschitzness. Subsection SM2.2 provides additional discussions on this matter. Additionally, subsection 3.1 offers theoretical intuitions for selecting mixing probabilities over mixing logits.

With these design choices implemented, the formulation (3.3) can be reparameterized as

$$(3.4) \quad h_i^\alpha(x) := \log \left((1 - \alpha) \sigma \circ g_i(x) + \alpha \cdot \sigma \circ h_i(x) \right) \quad \forall i \in [c],$$

where $\alpha = \frac{\gamma}{1+\gamma} \in [0, 1]$. We take $h^\alpha(\cdot)$ in (3.4), which is a convex combination of base classifier probabilities, as our proposed mixed classifier. Note that (3.4) calculates the mixed classifier logits, acting as a drop-in replacement for existing models which usually produce logits. Removing the logarithm recovers the output probabilities without changing the predicted class.

3.1. Theoretical certified robust radius. In this section, we derive certified robust radii for the mixed classifier $h^\alpha(\cdot)$ introduced in (3.4), given in terms of the robustness properties of $h(\cdot)$ and the mixing parameter α . The results ensure that despite being more sophisticated than a single model, $h^\alpha(\cdot)$ cannot be easily conquered, even if an adversary attempts to adapt its attack methods to its structure. Such guarantees are of paramount importance for reliable deployment in safety-critical applications. Note that while the focus of this paper is improved empirical accuracy-robustness trade-off and the existing literature often considers empirical and certified robustness separately, we will discuss how the certified results in this section provide important insights into the empirical performance, as the underlying assumptions are realistic and (approximately) verifiable for many empirically robust models.

Noticing that the base model probabilities satisfy $0 \leq \sigma \circ g_i(\cdot) \leq 1$ and $0 \leq \sigma \circ h_i(\cdot) \leq 1$ for all i , we introduce the following generalized and tightened notion of certified robustness.

Definition 3.1. Consider a model $h : \mathbb{R}^d \rightarrow \mathbb{R}^c$ and an arbitrary input $x \in \mathbb{R}^d$. Further consider $y = \arg \max_i h_i(x)$, $\mu \in [0, 1]$, and $r \geq 0$. Then, $h(\cdot)$ is said to be certifiably robust at x with margin μ and radius r if $\sigma \circ h_y(x + \delta) \geq \sigma \circ h_i(x + \delta) + \mu$ for all $i \neq y$ and all $\delta \in \mathbb{R}^d$ such that $\|\delta\|_p \leq r$.

Intuitively, Definition 3.1 ensures that all points within a radius from a nominal point have the same prediction as the nominal point, with the difference between the top and runner-up probabilities no smaller than a threshold. For practical classifiers, the robust margin can be straightforwardly estimated by calculating the confidence gap between the predicted and the runner-up classes at an adversarial input obtained with strong attacks. As shown in the experiments in subsection 5.1.2, if a real-world robust model is robust at some input with a given radius, it is likely to be robust with a nontrivial margin.

Lemma 3.2. *Let $x \in \mathbb{R}^d$ and $r \geq 0$. If it holds that $\alpha \in [\frac{1}{2}, 1]$ and $h(\cdot)$ is certifiably robust at x with margin $\frac{1-\alpha}{\alpha}$ and radius r , then the mixed classifier $h^\alpha(\cdot)$ is robust in the sense that $\arg \max_i h_i^\alpha(x + \delta) = \arg \max_i h_i(x)$ for all $\delta \in \mathbb{R}^d$ such that $\|\delta\|_p \leq r$.*

Proof. Suppose that $h(\cdot)$ is certifiably robust at x with margin $\frac{1-\alpha}{\alpha}$ and radius r . Since $\alpha \in [\frac{1}{2}, 1]$, it holds that $\frac{1-\alpha}{\alpha} \in [0, 1]$. Let $y = \arg \max_i h_i(x)$. Consider an arbitrary $i \in [c] \setminus \{y\}$ and $\delta \in \mathbb{R}^d$ such that $\|\delta\|_p \leq r$. Since $\sigma \circ g_i(x + \delta) \in [0, 1]$, it holds that

$$\begin{aligned} \exp(h_y^\alpha(x + \delta)) - \exp(h_i^\alpha(x + \delta)) &= (1 - \alpha)(\sigma \circ g_y(x + \delta) - \sigma \circ g_i(x + \delta)) + \alpha(\sigma \circ h_y(x + \delta) - \sigma \circ h_i(x + \delta)) \\ &\geq (1 - \alpha)(0 - 1) + \alpha(\sigma \circ h_y(x + \delta) - \sigma \circ h_i(x + \delta)) \\ &\geq (\alpha - 1) + \alpha\left(\frac{1-\alpha}{\alpha}\right) = 0. \end{aligned}$$

Thus, it holds that $h_y^\alpha(x + \delta) \geq h_i^\alpha(x + \delta)$ for all $i \neq y$, and thus $\arg \max_i h_i^\alpha(x + \delta) = y = \arg \max_i h_i(x)$. ■

While most existing provably robust results consider the special case with zero margin, we will show that models built via common methods are also robust with nonzero margins. We specifically consider two types of popular robust classifiers: Lipschitz continuous models (Theorem 3.5) and RS models (Theorem SM1.2). Here, Lemma 3.2 builds the foundation for proving these two theorems, which amounts to showing that Lipschitz and RS models are robust with nonzero margins and thus the mixed classifiers built with them are robust. Lemma 3.2 can also motivate future researchers to develop margin-based robustness guarantees for base classifiers so that they immediately grant robustness guarantees for mixed architectures.

Lemma 3.2 additionally provides further justifications for using probabilities instead of logits in the smoothing operation. Intuitively, $(1 - \alpha)\sigma \circ g_i(\cdot)$ is bounded between 0 and $1 - \alpha$, so as long as α is relatively large (specifically, at least $\frac{1}{2}$), the detrimental effect of $g(\cdot)$ when subject to attack can be overcome by $h(\cdot)$. Had we used the logits $g_i(\cdot)$, since this quantity cannot be bounded, it would have been much harder to overcome the vulnerability of $g(\cdot)$.

Since we do not make assumptions on the Lipschitzness or robustness of $g(\cdot)$, Lemma 3.2 is tight. To understand this, we suppose that there exist some $i \in [c] \setminus \{y\}$ and $\delta \neq 0$ such that $\|\delta\|_p \leq r$ that make $\sigma \circ h_y(x + \delta) - \sigma \circ h_i(x + \delta) := h_d$ smaller than $\frac{1-\alpha}{\alpha}$, indicating that $-\alpha h_d > \alpha - 1$. Since the only information about $g(\cdot)$ is that $\sigma \circ g_i(x + \delta) \in [0, 1]$ and thus the value $\sigma \circ g_y(x + \delta) - \sigma \circ g_i(x + \delta) := g_d$ can be any number between -1 and 1 , it is possible that $(1 - \alpha)g_d$ is smaller than $-\alpha h_d$. By (3.4), when $(1 - \alpha)g_d < -\alpha h_d$, it holds that $h_y^\alpha(x + \delta) < h_i^\alpha(x + \delta)$, and thus $\arg \max_i h_i^\alpha(x + \delta) \neq \arg \max_i h_i(x)$.

Definition 3.3. *A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is called ℓ_p -Lipschitz continuous if there exists $L \in (0, \infty)$ such that $|f(x') - f(x)| \leq L\|x' - x\|_p$ for all $x', x \in \mathbb{R}^d$. The Lipschitz constant of such f is defined to be*

$$\text{Lip}_p(f) := \inf \left\{ L \in (0, \infty) : |f(x') - f(x)| \leq L\|x' - x\|_p \ \forall x', x \in \mathbb{R}^d \right\}.$$

Assumption 3.4. *The base model $h(\cdot)$ is robust in the sense that, for all $i \in \{1, 2, \dots, n\}$, $\sigma \circ h_i(\cdot)$ is ℓ_p -Lipschitz continuous with Lipschitz constant $\text{Lip}_p(\sigma \circ h_i)$.*

Theorem 3.5. Suppose that Assumption 3.4 holds, and let $y = \arg \max_i h_i(x)$, where $x \in \mathbb{R}^d$ is arbitrary. Then, if $\alpha \in [\frac{1}{2}, 1]$, it holds that $\arg \max_i h_i^\alpha(x + \delta) = y$ for all $\delta \in \mathbb{R}^d$ such that

$$(3.5) \quad \|\delta\|_p \leq r_{\text{Lip},p}^\alpha(x) := \min_{i \neq y} \frac{\alpha(\sigma \circ h_y(x) - \sigma \circ h_i(x)) + \alpha - 1}{\alpha(\text{Lip}_p(\sigma \circ h_y) + \text{Lip}_p(\sigma \circ h_i))}.$$

Proof. Suppose that $\alpha \in [\frac{1}{2}, 1]$, and let $\delta \in \mathbb{R}^d$ be such that $\|\delta\|_p \leq r_{\text{Lip},p}^\alpha(x)$. Furthermore, let $i \in [c] \setminus \{y\}$. It holds that

$$\begin{aligned} \sigma \circ h_y(x + \delta) - \sigma \circ h_i(x + \delta) &= \sigma \circ h_y(x) - \sigma \circ h_i(x) + \sigma \circ h_y(x + \delta) - \sigma \circ h_y(x) + \sigma \circ h_i(x) - \sigma \circ h_i(x + \delta) \\ &\geq \sigma \circ h_y(x) - \sigma \circ h_i(x) - \text{Lip}_p(\sigma \circ h_y) \|\delta\|_p - \text{Lip}_p(\sigma \circ h_i) \|\delta\|_p \\ &\geq \sigma \circ h_y(x) - \sigma \circ h_i(x) - (\text{Lip}_p(\sigma \circ h_y) + \text{Lip}_p(\sigma \circ h_i)) r_{\text{Lip},p}^\alpha(x) \geq \frac{1-\alpha}{\alpha}. \end{aligned}$$

Therefore, $h(\cdot)$ is certifiably robust at x with margin $\frac{1-\alpha}{\alpha}$ and radius $r_{\text{Lip},p}^\alpha(x)$. Hence, by Lemma 3.2, the claim holds. ■

Note that the ℓ_p norm that we certify can be arbitrary (e.g., ℓ_1 , ℓ_2 , or ℓ_∞), so long as the Lipschitz constant of the robust network $h(\cdot)$ is computed with respect to the same norm.

Assumption 3.4 is not restrictive in practice. For example, Gaussian RS with smoothing variance $\sigma^2 I_d$ (I_d is the identity matrix in $\mathbb{R}^{d \times d}$) yields robust models with ℓ_2 -Lipschitz constant $\sqrt{2/\pi\sigma^2}$ [75]. In subsection SM1.3, we use experiments to verify the certified robustness of our method when $h(\cdot)$ is an RS model. Additionally, methods have been proposed to compute upper bounds on neural network Lipschitz constants, thus allowing our certified robustness guarantees via Assumption 3.4 and Theorem 3.5 to be employed [35, 48, 79]. The notion of Lipschitz continuity has even motivated novel robustness methods [64, 80, 70].

Assumption 3.4 can be relaxed to the even less restrictive scenario of using local Lipschitz constants over a neighborhood (e.g., a norm ball) around a nominal input x (i.e., how flat $\sigma \circ h(\cdot)$ is near x) as a surrogate for the global Lipschitz constants. In this case, Theorem 3.5 holds for all δ within this neighborhood. Specifically, suppose that for an arbitrary input x and an ℓ_p attack radius ϵ , it holds that $\sigma \circ h_y(x) - \sigma \circ h_y(x + \delta) \leq \epsilon \cdot \text{Lip}_p^x(\sigma \circ h_y)$ and $\sigma \circ h_i(x + \delta) - \sigma \circ h_i(x) \leq \epsilon \cdot \text{Lip}_p^x(\sigma \circ h_i)$ for all $i \neq y$ and all perturbations δ such that $\|\delta\|_p \leq \epsilon$. Furthermore, suppose that the robust radius $r_{\text{Lip},p}^\alpha(x)$, as defined in (3.5) but using the local Lipschitz constant Lip_p^x as a surrogate to the global constant Lip_p , is not smaller than ϵ . Then, if the robust base classifier $h(\cdot)$ is correct at the nominal point x , then the mixed classifier $h^\alpha(\cdot)$ is robust at x within the radius ϵ . The proof follows that of Theorem 3.5.

The relaxed Lipschitzness defined above can be estimated for practical differentiable classifiers via an algorithm derived from the PGD attack [87]. The authors of [87] showed that many existing empirically robust models, including those trained with AT or TRADES, are locally Lipschitz. Note that [87] evaluates the local Lipschitz constants of the logits, whereas we analyze the probabilities, whose Lipschitz constants are much smaller, and small enough to certify a meaningful robust radius. Hence, Theorem 3.5 provides important insights into the empirical robustness of the mixed classifier. A detailed discussion is presented in subsection SM3.3.

An intuitive explanation of Theorem 3.5 is that if α approaches 1, then $r_{\text{Lip},p}^\alpha(x)$ approaches $\min_{i \neq y} \frac{h_y(x) - h_i(x)}{\text{Lip}_p(h_y) + \text{Lip}_p(h_i)}$, which is the standard (global) Lipschitz-based robust radius of $h(\cdot)$ around x (see, e.g., [35, 43] for further discussions on Lipschitz-based robustness). On the other hand, if α is too small compared to the relative confidence of $h(\cdot)$, namely, if there exists $i \neq y$ such that $\alpha \leq \frac{1}{1 + \sigma \circ h_y(x) - \sigma \circ h_i(x)}$, then $r_{\text{Lip},p}^\alpha(x)$ is nonpositive, and in this case we cannot provide nontrivial certified robustness for $h^\alpha(\cdot)$. This is rooted in the fact that too small of an α value amounts to excess weight in the nonrobust classifier $g(\cdot)$. If $h(\cdot)$ is 100% confident in its prediction, then $\sigma \circ h_y(x) - \sigma \circ h_i(x) = 1$ for all $i \neq y$, and therefore this threshold value of α becomes $\frac{1}{2}$, leading to nontrivial certified radii for $\alpha > \frac{1}{2}$. However, once we put over $\frac{1}{2}$ of the weight into $g(\cdot)$, a nonzero radius around x is no longer certifiable. Since there are no assumptions on the robustness of $g(\cdot)$ around x , this is intuitively the best one can expect.

To summarize our certified robustness results, Lemma 3.2 shows the connection between the robust margin of the robust classifier and the robustness of the mixed classifier, while Theorem 3.5 demonstrates how general Lipschitz robust base classifiers exploit this relationship. Since empirically robust models often satisfy the conditions of these two results, they guarantee that adaptive attacks cannot easily circumvent our proposed robustification.

In subsection SM1.1 in the supplementary material, we further tighten the certified radius estimation in the special case when $h(\cdot)$ is a randomized smoothing classifier and the robust radius is defined with the ℓ_2 norm. We achieve this by exploiting the stronger Lipschitzness of $x \mapsto \Phi^{-1}(\sigma \circ h_i(x))$ arising from the unique structure granted by Gaussian convolution operations (Φ^{-1} is the inverse Gaussian cumulative distribution function). In subsection SM1.3, we compare the certified robustness of the mixed classifier to existing certifiably robust methods.

4. Adaptive smoothing strength with the mixing network. So far, α has been treated as a fixed hyperparameter. A more intelligent approach is to allow α to be different for each x by using a function $\alpha(x)$. We take $\alpha(x)$ to be deterministic, as stochastic defenses are challenging to properly evaluate.

One motivation for adopting the adaptive mixing ratio $\alpha(x)$ is that the optimal α^* varies when x changes. For example, when x is unperturbed, the standard model $g(\cdot)$ outperforms the robust base model $h(\cdot)$. If x is an attacked input targeting $g(\cdot)$, then $h(\cdot)$ should again be used. However, if the attack target is $h(\cdot)$, then as shown in Figure 2, even though $h(\cdot)$ is robust, feeding x into $g(\cdot)$ is a better choice. This is because the vulnerabilities of $g(\cdot)$ and $h(\cdot)$ differ enough that an adversarial perturb targeting $h(\cdot)$ is benign to $g(\cdot)$.

When the adversary targets a mixed classifier $h^{\alpha_t}(\cdot)$, as α_t varies, the optimal strategy changes. Figure 2 provides a visualization based on the CIFAR-10 dataset. Specifically, we assemble a composite model $h^{\alpha_t}(\cdot)$ using a ResNet-18 standard classifier $g(\cdot)$ and a ResNet-18 robust classifier $h(\cdot)$ (both from [65]) via (3.4). Then, we attack $h^{\alpha_t}(\cdot)$ with different values of α_t via PGD₂₀, save the adversarial instances, and report the accuracy of $g(\cdot)$ and $h(\cdot)$ on these instances. When $\alpha_t \leq \text{Sigmoid}(5.72) = 0.9967$, the robust model $h(\cdot)$ performs better. When $\alpha_t > 0.9967$, the standard model $g(\cdot)$ is more suitable.

Throughout the remainder of this section, we overload the notation $h^\alpha(\cdot)$ even though $\alpha(\cdot)$ may be a function of the input, i.e., we define $h^\alpha(x) = h^{\alpha(x)}(x)$.

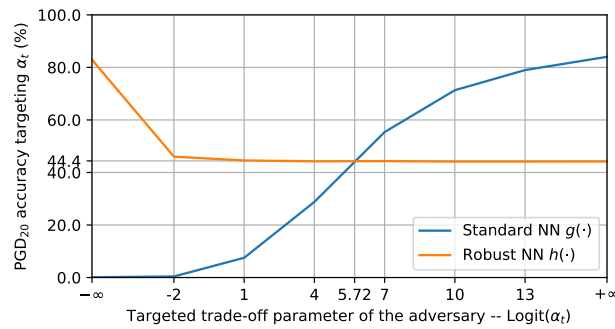


Figure 2. Attacked accuracy of the accurate base classifier $g(\cdot)$ and the robust base model $h(\cdot)$ when the adversary targets different values of α_t . For better readability, we use $\text{Logit}(\alpha_t)$ as the horizontal axis labels, where $\text{Logit}(\cdot)$ denotes the inverse function of Sigmoid.

4.1. The existence of $\alpha(x)$ that achieves the trade-off. The following theorem shows that, under realistic conditions, there exists a function $\alpha(\cdot)$ that makes the combined classifier correct whenever either $g(\cdot)$ and $h(\cdot)$ makes the correct prediction, which further implies that the combined classifier matches the clean accuracy of $g(\cdot)$ and the attacked accuracy of $h(\cdot)$.

Theorem 4.1. Let $\epsilon > 0$, $(x_1, y_1), (x_2, y_2) \sim \mathcal{D}$, and $y_1 \neq y_2$ (i.e., each input corresponds to a unique true label). Assume that $h_i(\cdot)$, $\|\nabla h_i(\cdot)\|_{p^*}$, and $\|\nabla g_i(\cdot)\|_{p^*}$ are all bounded and that there does not exist $z \in \mathbb{R}^d$ such that $\|z - x_1\|_p \leq \epsilon$ and $\|z - x_2\|_p \leq \epsilon$. Then, there exists a function $\alpha(\cdot)$ such that the assembled classifier $h^\alpha(\cdot)$ satisfies

$$\mathbb{P}_{(x,y) \sim \mathcal{D}} \left[\arg \max_{i \in [c]} h_i^\alpha(x + \delta) = y \right] \geq \max \left\{ \mathbb{P}_{(x,y) \sim \mathcal{D}, \delta \sim \mathcal{F}} [\arg \max_{i \in [c]} g_i(x + \delta) = y], \mathbb{P}_{(x,y) \sim \mathcal{D}, \delta \sim \mathcal{F}} [\arg \max_{i \in [c]} h_i(x + \delta) = y] \right\},$$

where \mathcal{F} is an arbitrary distribution that satisfies $\mathbb{P}_{\delta \sim \mathcal{F}}[\|\delta\|_p > \epsilon] = 0$.

Proof. Since it is assumed that the perturbation balls of the data are nonoverlapping, the true label y corresponding to each perturbed data $x + \delta$ with the property $\|\delta\|_p \leq \epsilon$ is unique. Therefore, the indicator function

$$\alpha(x + \delta) = \begin{cases} 0 & \text{if } \arg \max_{i \in [c]} g_i(x + \delta) = y, \\ 1 & \text{otherwise} \end{cases}$$

satisfies that

$$\alpha(x + \delta) = 1 \quad \text{if} \quad \arg \max_{i \in [c]} g_i(x + \delta) \neq y \quad \text{and} \quad \arg \max_{i \in [c]} h_i(x + \delta) = y.$$

Therefore, it holds that

$$\begin{aligned} h_i^\alpha(x + \delta) &= g_i(x + \delta) & \text{if} & \quad \arg \max_{i \in [c]} g_i(x + \delta) = y, \\ h_i^\alpha(x + \delta) &= h_i(x + \delta) & \text{if} & \quad \arg \max_{i \in [c]} g_i(x + \delta) \neq y \quad \text{and} \quad \arg \max_{i \in [c]} h_i(x + \delta) = y, \end{aligned}$$

implying that

$$\arg \max_{i \in [c]} h_i^\alpha(x + \delta) = y \quad \text{if} \quad \left(\arg \max_{i \in [c]} g_i(x + \delta) = y \quad \text{or} \quad \arg \max_{i \in [c]} h_i(x + \delta) = y \right),$$

which leads to the desired statement. ■

Note that the distribution \mathcal{F} is arbitrary, implying that the test data can be clean data, any type of adversarial data, or some combination of both. As a special case, when \mathcal{F} is a Dirac measure at the origin, Theorem 4.1 implies that the clean accuracy of $h^\alpha(\cdot)$ is as good as the standard classifier $g(\cdot)$. Conversely, when \mathcal{F} is a Dirac measure at the worst-case perturbation, the adversarial accuracy of $h^\alpha(\cdot)$ is not worse than the robust model $h(\cdot)$, implying that if $h(\cdot)$ is inherently robust, then $h^\alpha(\cdot)$ inherits the robustness. One can then conclude that there exists a $h^\alpha(\cdot)$ that matches the clean accuracy of $g(\cdot)$ and the robustness of $h(\cdot)$.

While Theorem 4.1 guarantees the existence of an instance of $\alpha(\cdot)$ that perfectly balances accuracy and robustness, finding an $\alpha(\cdot)$ that achieves this trade-off can be hard. However, we will use experiments to show that an $\alpha(\cdot)$ represented by a neural network can retain most of the robustness of $h(\cdot)$ while greatly boosting the clean accuracy. In particular, while we used the case of $\alpha(\cdot)$ being an indicator function to demonstrate the possibility of achieving the trade-off, Figure 1 has shown that letting α take an appropriate value between 0 and 1 also improves the trade-off. Thus, the task for the neural approximator is easier than representing the indicator function. Also note that if certified robustness is desired, one can enforce a lower bound on $\alpha(\cdot)$ and take advantage of Theorem 3.5 while still enjoying the mitigated trade-off.

4.2. Attacking the adaptive classifier. When the combined model $h^\alpha(\cdot)$ is under adversarial attack, the function $\alpha(\cdot)$ provides an additional gradient flow path. Intuitively, the attack should be able to force α to be small through this additional gradient path, tricking the mixing network into favoring the nonrobust $g(\cdot)$. Following the guidelines for constructing adaptive attacks [81], in the experiments, we consider the following types of attacks:

- A Gray-box PGD₂₀:** The adversary has access to the gradients of $g(\cdot)$ and $h(\cdot)$ when performing first-order optimization, but is not given the gradient of the mixing network $\alpha(\cdot)$. We consider untargeted PGD attack with a fixed initialization.
- B White-box PGD₂₀:** Since the mixed classifier is end-to-end differentiable, we follow [81] and allow the attack to query end-to-end gradient, including that of the mixing network.
- C White-box AutoAttack:** AutoAttack is a stronger attack formed by an ensemble of four attack algorithms [31]. The method considers Auto-PGD (APGD) attacks with the untargeted cross-entropy loss and the targeted Difference of Logits Ratio loss, in addition to the targeted FAB attack and the black-box Square Attack (SA) [10]. Again, the end-to-end mixed classifier gradient is available to the adversary. AutoAttack requires much more computation than PGD₂₀.
- D Adaptive white-box AutoAttack:** Since the mixing network is a crucial component of the defense, we add an APGD loss component that aims to decrease α into AutoAttack to specifically target the mixing network.

We will show that the adaptively smoothed model is robust against the attack that it is trained against. When trained using untargeted and targeted APGD₇₅ attacks, our model becomes robust against AutoAttack while noticeably improving the accuracy-robustness trade-off. In subsection 5.1.1, we additionally consider evaluating with transfer attacks.

4.3. The mixing network. In practice, we use a neural network $\alpha_\theta(\cdot) : \mathbb{R}^d \rightarrow [0, 1]$ to learn an effective mixing network that adjusts the outputs of $g(\cdot)$ and $h(\cdot)$. Here, θ represents the

trainable parameters of the mixing network, and we refer to $\alpha_\theta(\cdot)$ as the “mixing network.” To enforce an output range constraint, we apply a Sigmoid function to the mixing network output. Note that when training the mixing network $\alpha_\theta(\cdot)$, the base classifiers $g(\cdot)$ and $h(\cdot)$ are frozen. Freezing the base classifiers allows the mixed classifier to take advantage of existing accurate models and their robust counterparts, maintaining explainability and avoiding unnecessary feature distortions that the adversary can potentially exploit.

The mixing network’s task of treating clean and attacked inputs differently is closely related to adversary detection. To this end, we adapt the detection architecture introduced in [62] for our mixing network. This architecture achieves high performance and low complexity and is end-to-end differentiable, enabling convenient training and evaluation. While the authors of [21] argued that simultaneously attacking the base classifier and the adversary detector can bring the detection rate of the detection method proposed in [62] to near zero, we show that with several key modifications, the method is effective even against strong white-box attacks. Specifically, our mixing network $\alpha_\theta(\cdot)$ takes advantage of both base models $g(\cdot)$ and $h(\cdot)$ by concatenating their intermediate features ([62] used only one base model). More importantly, we include stronger adaptive adversaries during training to generate much more diverse training examples.

The mixing network structure is based on a ResNet-18, which is known to perform well for a wide range of computer vision applications and is often considered the go-to architecture. We make some minimal necessary changes to ResNet-18 for it to fit into our framework. Specifically, as the mixing network takes information from both $g(\cdot)$ and $h(\cdot)$, it uses the concatenated embeddings from the base classifiers. While [62] considers a single ResNet as the base classifier and uses the embeddings after the first ResNet block, to avoid the potential vulnerability against “feature adversaries” [74], we consider the embeddings from two different layers of the base model. Figure 3 demonstrates the modified architecture. The detailed implementations used in the experiment section are discussed in subsection SM4.1.

Since Figure 1 shows that even a constant α can alleviate the accuracy-robustness trade-off, our method does not excessively rely on the performance of the mixing network $\alpha_\theta(\cdot)$. In subsection 5.2, we provide empirical results demonstrating that the above modifications help the overall mixed network defend against strong attacks.

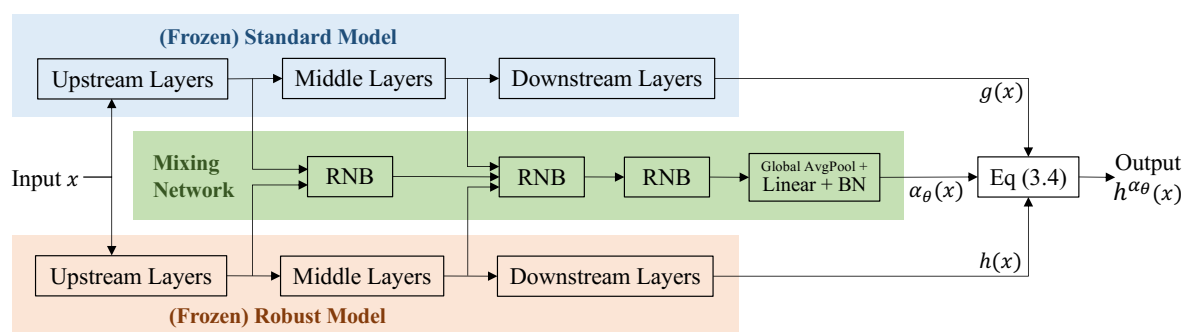


Figure 3. The overall architecture of the adaptively smoothed classifier introduced in section 4. “RNB” stands for ResNetBlock and “BN” represents the 2D batch normalization layer.

4.4. Training the mixing network. Consider the following two loss functions for training the mixing network $\alpha_\theta(\cdot)$:

- **Multiclass cross-entropy:** We minimize the multiclass cross-entropy loss of the combined classifier, which is the ultimate goal of the mixing network:

$$(4.1) \quad \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\ell_{\text{CE}}(h^{\alpha_\theta}(x + \delta), y) \right],$$

where ℓ_{CE} is the cross-entropy (CE) loss for logits and $y \in [c]$ is the label corresponding to x . The base classifiers $g(\cdot)$ and $h(\cdot)$ are frozen and not updated. Again, δ denotes the perturbation, and the distribution \mathcal{F} is arbitrary. In our experiments, to avoid overfitting to a particular attack radius, \mathcal{F} is formed by perturbations with randomized radii.

- **Binary cross-entropy:** The optimal α^* that minimizes ℓ_{CE} in (4.1) can be estimated for each training point. Specifically, depending on whether the input is attacked and how it is attacked, either $g(\cdot)$ or $h(\cdot)$ should be prioritized. Thus, we treat the task as a binary classification problem and solve the optimization problem

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\ell_{\text{BCE}}(\alpha_\theta(x + \delta), \tilde{\alpha}) \right],$$

where ℓ_{BCE} is the binary cross-entropy (BCE) loss for probabilities and $\tilde{\alpha} \in \{0, 1\}$ is the “pseudolabel” for the output of the mixing network that approximates α^* .

Using only the multiclass loss suffers from a distribution mismatch between training and test data. Specifically, the robust classifier $h(\cdot)$ may achieve a low loss on adversarial training data but a high loss on test data. For example, with our ResNet-18 robust CIFAR-10 classifier, the PGD₁₀ adversarial training and test accuracy are very different, at 93.01% and 45.55%, respectively. As a result, approximating (4.1) with empirical risk minimization on training data does not effectively optimize the true risk. To understand this, notice that when the adversary perturbs a test input x targeting $h(\cdot)$, the standard classifier prediction $g(x)$ yields a lower loss than $h(x)$. However, if x is an attacked example in the training set, then $g(x)$ and $h(x)$ have similar losses, and the mixing network does not receive an incentive to choose $g(\cdot)$ when detecting an attack targeting $h(\cdot)$.

The binary loss, on the other hand, does not capture the potentially different sensitivity of each input. Certain inputs can be more vulnerable to adversarial attacks, and ensuring the correctness of the mixing network on these inputs is more crucial.

To this end, we combine the above two components into a composite loss function, incentivizing the mixing network to select the standard classifier $g(\cdot)$ when appropriate, while forcing it to remain conservative. The composite loss for each data-label pair (x, y) is

$$(4.2) \quad \begin{aligned} \ell_{\text{composite}}(\theta, (x, y, \tilde{\alpha})) = & c_{\text{CE}} \cdot \ell_{\text{CE}}(h^{\alpha_\theta}(x + \delta), y) + c_{\text{BCE}} \cdot \ell_{\text{BCE}}(\alpha_\theta(x + \delta), \tilde{\alpha}) \\ & + c_{\text{prod}} \cdot \ell_{\text{CE}}(h^{\alpha_\theta}(x + \delta), y) \cdot \ell_{\text{BCE}}(\alpha_\theta(x + \delta), \tilde{\alpha}), \end{aligned}$$

where the hyperparameters c_{CE} , c_{BCE} , and c_{prod} control the weights of the loss components. Subsection SM3.2 in the supplementary material discusses how these hyperparameters affect the performance of the trained mixing model.

5. Numerical experiments.

5.1. Robust neural network smoothing with a fixed strength. We first consider the case where the smoothing strength α is a fixed value. In this section, we focus on using empirically robust base classifiers and consider the CIFAR-10 dataset. In subsection SM1.3 in the supplementary material, we present the certified robustness results when the robust base model is based on RS, simultaneously instantiating Lemma 3.2 and Theorem 3.5. In subsection SM3.3, we show that empirically robust models can also take advantage of our theoretical analyses by estimating their Lipschitz constant.

5.1.1. α 's influence on mixed classifier robustness. We first analyze how the accuracy of the mixed classifier changes with the mixing strength α under various settings. Specifically, we consider PGD₂₀ attacks that target $g(\cdot)$ and $h(\cdot)$ individually (denoted as STD and ROB attacks), in addition to the adaptive PGD₂₀ attack generated using the end-to-end gradient of $h^\alpha(\cdot)$, denoted as the MIX attack. Note that the STD and ROB attacks, which share the inspiration of [88], correspond to the “transfer attack” setting, a common black-box attack strategy designed for defenses with unavailable or unreliable gradients. Note that the models with the best transferability with the mixed classifier $h^\alpha(\cdot)$ would likely be its base classifiers $g(\cdot)$ and $h(\cdot)$, precisely corresponding to the STD and ROB attack settings.

We use a ResNet18 model trained on clean data as the standard base classifier $g(\cdot)$ and use another ResNet18 trained on PGD₂₀ data as the robust base classifier $h(\cdot)$. The test accuracy corresponding to each α value is presented in Figure 4. As α increases, the clean accuracy of $h^\alpha(\cdot)$ converges from the clean accuracy of $g(\cdot)$ to the clean accuracy of $h(\cdot)$. In terms of the attacked performance, when the attack targets $g(\cdot)$, the attacked accuracy increases with α . When the attack targets $h(\cdot)$, the attacked accuracy decreases with α , showing that the attack targeting $h(\cdot)$ becomes more benign when the mixed classifier emphasizes $g(\cdot)$. When the attack targets $h^\alpha(\cdot)$, the attacked accuracy increases with α .

When α is around 0.5, the MIX-attacked accuracy of $h^\alpha(\cdot)$ quickly increases from near zero to more than 30% (which is two-thirds of $h(\cdot)$'s attacked accuracy). This observation

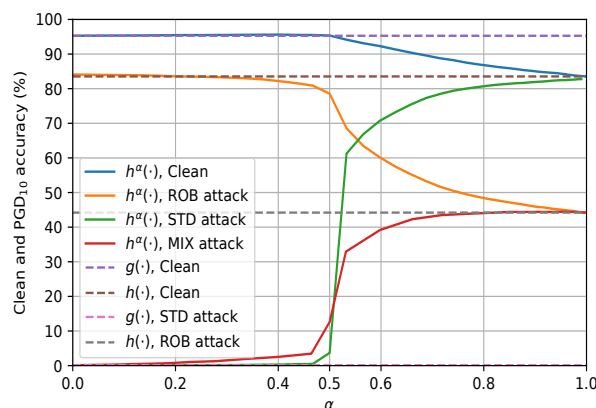


Figure 4. The performance of the mixed classifier $h^\alpha(\cdot)$. “STD attack,” “ROB attack,” and “MIX attack” refer to the PGD₂₀ attack generated using the gradient of $g(\cdot)$, $h(\cdot)$, and $h^\alpha(\cdot)$, respectively, with ϵ set to $8/255$.

precisely matches the theoretical intuition provided by Theorem 3.5. When α is greater than 0.5, the clean accuracy gradually decreases at a much slower rate, leading to the noticeably alleviated accuracy-robustness trade-off. Note that this improved trade-off is achieved without any further training beyond the weights of $g(\cdot)$ and $h(\cdot)$. When α is greater than 0.55, neither STD attack nor ROB attack can reduce the accuracy of the mixed classifier below the end-to-end gradient-based attack (MIX attack), indicating that the considered transfer attack is weaker than gradient-based attack for practical α values, and implying that the robustness of $h^\alpha(\cdot)$ does not rely on obfuscated gradients. In subsection 5.1.2, we will reveal that the source of $h^\alpha(\cdot)$'s robustness lies in $h(\cdot)$'s well-calibrated confidence properties.

5.1.2. The relationship between $h^\alpha(\cdot)$'s robustness and $h(\cdot)$'s confidence. Our theoretical analysis (Lemma 3.2) has highlighted the relationship between the mixed classifier robustness and the robust base classifier $h(\cdot)$'s robust margin. For practical models, the margin at a given radius can be estimated with the confidence gap between the predicted and runner-up classes evaluated on strongly adversarial inputs, such as images returned from PGD₂₀ or AutoAttack. Moreover, the improved accuracy-robustness trade-off of the mixed classifier, as evidenced by the difference in how the clean and the attacked accuracy change with α in Figure 4, can also be explained by the prediction confidence of $h(\cdot)$.

According to Table 1, the robust base classifier $h(\cdot)$ makes confident correct predictions even when under attack (average robust margin is 0.768 evaluated with PGD₂₀ and 0.774 with AutoAttack¹). Moreover, the robust margin of $h(\cdot)$ follows a long-tail distribution. Specifically, the median robust margin is 0.933 (same number when evaluated with PGD₂₀ or AutoAttack), much larger than the 0.768/0.774 average margin. Thus, most attacked inputs correctly classified by $h(\cdot)$ are highly confident (i.e., robust with large margins), with only a tiny portion suffering from small robust margins. As Lemma 3.2 suggests, such a property is precisely what adaptive smoothing relies on. Intuitively, once α becomes greater than 0.5 and gives $h(\cdot)$ more authority over $g(\cdot)$, $h(\cdot)$ can use its high confidence to correct $g(\cdot)$'s mistakes under attack.

On the other hand, $h(\cdot)$ is unconfident when it produces incorrect predictions on unattacked clean data, with the top two classes' output probabilities separated by merely 0.434. This probability gap again forms a long-tail distribution (the median is 0.378 which is less than

Table 1

Average gap between the probabilities of the predicted class and the runner-up class. $g(\cdot)$ and $h(\cdot)$ are the same ones used in Figure 4. The confidence difference highlighted by the bold numbers is crucial to the mitigated accuracy-robustness trade-off of the mixed classifier.

	Clean			PGD ₂₀			AutoAttack		
	Accuracy	✓ Gap	✗ Gap	Accuracy	✓ Gap	✗ Gap	Accuracy	✓ Gap	✗ Gap
$g(\cdot)$	95.28%	0.982	0.698	0.10%	0.602	0.998	0.00%	—	0.986
$h(\cdot)$	83.53%	0.854	0.434	44.17%	0.768	0.635	40.75%	0.774	0.553

✓ Gap: The average gap between the confidences of the predicted class and the runner-up class among all correctly predicted validation data.

✗ Gap: The same quantity evaluated among all incorrectly predicted validation data.

¹The calculation details the AutoAttacked confidence gap are presented in subsection SM4.2 of the supplementary material.

the mean), confirming that $h(\cdot)$ is generally unconfident when mispredicting and rarely makes confident incorrect predictions. Now, consider clean data that $g(\cdot)$ correctly classifies and $h(\cdot)$ mispredicts. Recall that we assume $g(\cdot)$ to be more accurate but less robust, so this scenario should be common. Since $g(\cdot)$ is confident (average top two classes' probability gap is 0.982) and $h(\cdot)$ is usually unconfident, even when $\alpha > 0.5$ and $g(\cdot)$ has less authority than $h(\cdot)$ in the mixture, $g(\cdot)$ can still correct some of the mistakes from $h(\cdot)$.

In summary, $h(\cdot)$ is confident when making correct predictions on attacked data, enjoying the large robust margin required by Lemma 3.2. At the same time, $h(\cdot)$ is unconfident when misclassifying clean data, and such a confidence property is the key source of the mixed classifier's improved accuracy-robustness trade-off. Additional analyses in subsection SM2.2 with alternative base models imply that multiple existing robust classifiers share the favorable confidence property and thus help the mixed classifier improve the trade-off.

The standard nonrobust classifier $g(\cdot)$ often does not have this desirable property: even though it is confident on clean data as are robust classifiers, it also makes highly confident mistakes under attack. Note that this does not undermine the mixed classifier robustness, since our formulation does not assume any robustness or smoothness from $g(\cdot)$.

5.1.3. Comparing the accuracy-robustness trade-off with existing methods. This subsection compares the accuracy-robustness trade-off of the mixed classifiers with existing baseline methods that emphasize addressing this trade-off.

TRADES [90] is one of the most famous and popular methods to improve the accuracy-robustness trade-off. Specifically, it trains robust models by minimizing the risk function

$$\mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\ell_{\text{CE}}(h(x), y) + \beta \max_{\|\delta\| \leq \epsilon} \ell_{\text{surrogate}}(h(x + \delta), h(x)) \right],$$

where $\beta \geq 0$ is a trade-off parameter between the two loss components and $\ell_{\text{surrogate}}$ is the “surrogate loss” that promotes robustness. The larger β is, the more robust the trained model becomes at the expense of clean accuracy. By adjusting β , we can adjust the accuracy-robustness trade-off of TRADES similarly to adjusting α in our mixed classifier.

The authors of [90] reported that $\beta = 6$ optimized the adversarial robustness and released the corresponding model. We use this model and train three additional models with β set to 0, 0.1, and 0.3. Here, $\beta = 0$ is standard training, and the other two numbers were chosen so that the model accuracy spreads relatively uniformly between $\beta = 0$ and $\beta = 6$. All TRADES models use the WideResNet-34-10 architecture as in [90]. For a fair comparison, we build mixed classifiers using the TRADES model trained with $\beta = 0$ as $g(\cdot)$ and the $\beta = 6$ model as $h(\cdot)$. We compare the relationship between the PGD₂₀ accuracy and the clean accuracy in Figure 5. Note that the trade-off curve of the mixed classifier intercepts the TRADES curve at the two ends (since the models are exactly the same at the two ends) and is significantly above the TRADES in the middle, indicating that the accuracy-robustness trade-off of the mixed classifier is much more benign than that of TRADES.

IAAT [18] and Properly Learned Smoothing (PLS) [25] are two additional high-performance methods for alleviating the accuracy-robustness trade-off. Specifically, IAAT uses input-dependent attack budgets during adversarial training, while PLS performs stochastic weight averaging and smooths the logits via knowledge distillation and self-training. IAAT and PLS do not explicitly allow for adjusting between clean accuracy and adversarial robustness.

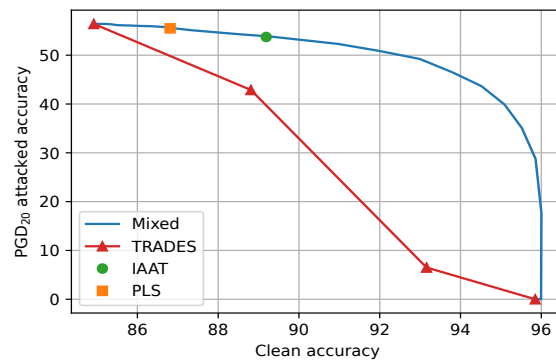


Figure 5. An accuracy-robustness trade-off comparison between our mixed classifier $h^\alpha(\cdot)$, denoted as “Mixed,” and TRADES, IAAT, and PLS models. TRADES allows for sweeping between accuracy and robustness, whereas IAAT and PLS are nonadjustable.

We implement IAAT on the same WideResNet-34-10 model architecture and add the result to Figure 5. For PLS, we use the accuracy reported in [25]. It can be observed that the TRADES-based mixed classifier achieves a similar accuracy-robustness trade-off as IAAT and PLS, while allowing for sweeping between accuracy and robustness conveniently unlike previous models. Note that for TRADES, adjusting the trade-off requires training a new model, which is costly. Meanwhile, IAAT and PLS do not allow for explicitly adjusting the trade-off altogether (hence shown as single points in Figure 5). In contrast, for our mixing classifier, the trade-off can be adjusted at inference time by simply tuning α and does not require retraining. Thus, our method is much more flexible and efficient while achieving a benign Pareto curve.

Even though the clean-robust accuracy curve of adaptive smoothing overlaps with that of IAAT at a single point (89.19% clean, 53.73% robust), adaptive smoothing still improves the overall accuracy-robustness trade-off. Specifically, on top of IAAT’s result, adaptive smoothing can further reduce the error rate by 31% while only sacrificing 6% of the robustness by achieving $\sim 50\%$ / $\sim 92.5\%$ robust/clean accuracy. In scenarios that are more sensitive to clean data performance, such a result makes adaptive smoothing more advantageous than IAAT, whose level of clean accuracy improvement is relatively limited.

Moreover, as discussed in section 1 and confirmed in subsection 5.2.2, our mixed classifier can easily incorporate existing innovations that improve clean accuracy or adversarial robustness, whereas fusing these innovations into training-based methods such as TRADES, IAAT, and PLS can be much more complicated. Also note that Figure 5 considers a constant α value, and adapting α for different input values further alleviates the trade-off. To provide experimental evidence, in Figure SM4 in subsection SM3.1, we add the mixed classifier results achieved with better base classifiers to the trade-off curve.

5.2. Robust neural network smoothing with adaptive strength. Having validated the effectiveness of the mixing formulation described in (3.4), we are now ready to incorporate the mixing network for adaptive smoothing strength. As in section 4, we denote the parameterized mixing network by $\alpha_\theta(\cdot)$, and slightly abuse notation by denoting the composite classifier with adaptive smoothing strength given by $h^{\alpha_\theta}(\cdot)$ by $h^{\alpha_\theta(x)}(\cdot)$, which is defined by $h^{\alpha_\theta(x)}(x) = h^{\alpha_\theta(x)}(x)$.

CIFAR-10 and CIFAR-100 are two of the most universal robustness evaluation datasets, and thus we use them to benchmark adaptive smoothing. We consider ℓ_∞ attacks and use the AdamW optimizer [50] for training the mixing network $\alpha_\theta(\cdot)$. The training data for $\alpha_\theta(\cdot)$ include clean images and the corresponding types of attacked images (attack settings A, B, and C presented in subsection 4.2). For setting C (AutoAttack), the training data only includes targeted and untargeted APGD attacks, with the other two AutoAttack components, FAB and Square, excluded during training in the interest of efficiency but included for evaluation. To alleviate overfitting, when generating training-time attacks, we randomize the attack radius and the number of steps, and add a randomly weighted binary cross-entropy component that aims to decrease the mixing network output to the attack objective (thereby tricking it into favoring $g(\cdot)$). Additionally, subsection SM4.1 discusses the details of implementing the architecture in Figure 3 for the ResNet base classifiers used in our experiments. Subsection SM3.2 conducts an ablation study on the hyperparameters in the composite loss function (4.2).

5.2.1. Ablation studies regarding attack settings. We first use smaller base classifiers to analyze the behavior of adaptive smoothing by exploring various training and attack settings. The performance of the base models and the assembled mixed classifier are summarized in Table 2, where each column represents the performance of one mixed classifier. The results show that the adaptive smoothing model can defend against the attacks on which the underlying mixing network is trained. Specifically, for the attack setting A (gray-box PGD), $h^{\alpha_\theta}(\cdot)$ is able to achieve the same level of PGD₂₀-attacked accuracy as $h(\cdot)$ while retaining a similar level of clean accuracy as $g(\cdot)$. For the setting B (white-box PGD), the attack is allowed to follow the gradient path provided by $\alpha_\theta(\cdot)$ and deliberately evade the part of the adversarial input space recognized by $\alpha_\theta(\cdot)$. While the training task becomes more challenging, the improvement in the accuracy-robustness trade-off is still substantial. Furthermore, the composite model can generalize to examples generated via the stronger AutoAttack. For the setting C (AutoAttack), the difficulty of the training problem further escalates. While the performance of $h^{\alpha_\theta}(\cdot)$ on clean data slightly decreases, the mixing network can offer a more vigorous defense against AutoAttack data, still improving the accuracy-robustness trade-off.

Table 2

CIFAR-10 results of adaptive smoothing models trained with three different settings.

CIFAR-10 base classifier performances					
Model	Architecture	Clean	PGD ₂₀	AutoAttack	
$g(\cdot)$ (accurate)	ResNet-18 (standard non-robust training)	95.28%	0.12%	0.00%	
$h(\cdot)$ (robust)	WideResNet-34-10 (TRADES model [90])	84.92%	57.16%	53.09%	

CIFAR-10 adaptive smoothing mixed classifier $h^{\alpha_\theta}(\cdot)$ performance					
Training setting \ Eval data	Clean	A	B	C	D (adaptive AutoAttack)
A (gray-box PGD ₂₀)	92.05%	57.22%	56.63%	40.04%	39.85%
B (white-box PGD ₂₀)	92.07%	57.25%	57.09%	40.02%	39.70%
C (white-box AutoAttack)	91.51%	56.30%	56.29%	42.78%	42.66%

Table 3

CIFAR-100 results of adaptive smoothing models trained with the three settings.

CIFAR-100 base classifier performances					
Model	Architecture	Clean	PGD ₂₀	AutoAttack	
$g(\cdot)$ (accurate)	ResNet-152 (based on BiT [52])	91.38%	0.14%	0.00%	
$h(\cdot)$ (robust)	WideResNet-70-16 (from [40])	69.17%	40.86%	36.98%	
CIFAR-100 adaptive smoothing mixed classifier $h^{\alpha_\theta}(\cdot)$ performance					
Training setting \ Eval data	Clean	A	B	C	D (adaptive AutoAttack)
A (gray-box PGD ₂₀)	83.99%	40.04%	30.59%	23.54%	23.78%
B (white-box PGD ₂₀)	83.96%	39.80%	34.48%	26.37%	26.17%
C (white-box AutoAttack)	80.90%	39.26%	38.92%	32.94%	32.80%

Table 3 repeats the above analyses on the CIFAR-100 dataset. The results confirm that adaptive smoothing achieves even more significant improvements on the CIFAR-100 dataset. Notably, even for the most challenging attack setting C, $h^{\alpha_\theta}(\cdot)$ correctly classifies 1173 additional clean images compared with $h(\cdot)$ (cutting the error rate by a third) while making only 404 additional incorrect predictions on AutoAttacked inputs (increasing the error rate by merely 6.4 relative percent). Such results show that $\alpha_\theta(\cdot)$ is capable of approximating a robust high-performance mixing network when trained with sufficiently diverse attacked data. The fact that $h^{\alpha_\theta}(\cdot)$ combines the clean accuracy of $g(\cdot)$ and the robustness of $h(\cdot)$ highlights that our method significantly improves the accuracy-robustness trade-off.

5.2.2. Comparisons against existing SOTA methods. In this section, we use Table 4 to show that when using SOTA base classifiers, adaptive smoothing noticeably improves the accuracy-robustness trade-off over existing methods.

Since the literature has regarded AutoAttack [31] as one of the most reliable robustness evaluation methods (weaker attacks such as PGD are known to be circumventable), we select AutoAttack-evaluated robust models as baselines. We highlight that these baseline models should not be treated as competitors, since advancements in building robust classifiers can be incorporated into our framework as $h(\cdot)$, helping adaptive smoothing perform even better.

For the accurate base classifier $g(\cdot)$, we fine-tune the BiT ResNet-152 checkpoint (from [51], pretrained on ImageNet-21k) on CIFAR-10 or CIFAR-100. Following the recipe from [51], our CIFAR-10 model achieves a 98.50% clean accuracy and our CIFAR-100 model achieves 91.38%.

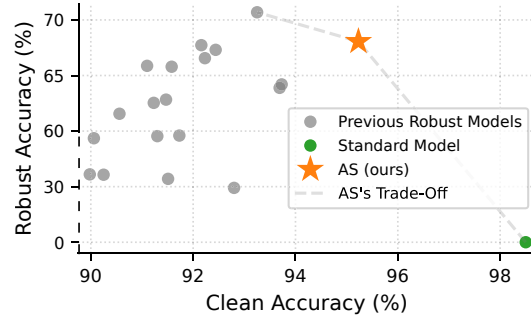
For CIFAR-10, we select the robust model checkpoint released in [86] as the robust base classifier $h(\cdot)$. Compared with $h(\cdot)$, adaptive smoothing retains 96.3 (relative) percent of the robust accuracy while reducing the clean data error rate by 29.3 (relative) percent. Among all models available on RobustBench as of this paper's submission, our method achieves the third highest AutoAttacked accuracy, behind only [86] (used as $h(\cdot)$ in our model) and [49] (for which AutoAttack is unreliable and the best-known attacked accuracy is lower than ours). Meanwhile, the clean accuracy of our mixed classifier is higher than all listed models with nontrivial ℓ_∞ robustness and is even higher than the listed nonrobust model that uses standard training.

Table 4

Clean and AutoAttack (AA) accuracy of adaptive smoothing (AS) compared with the reported accuracy of previous models. AS clearly improves the accuracy-robustness trade-off.

CIFAR-10

Method	Clean	AA
AS (adaptive smoothing, ours) *	95.23%	68.06%
SODEF+TRADES [50]	93.73%	71.28% [†]
Diffusion (EDM)+TRADES [87]	93.25%	70.69%
Diffusion (DDPM)+TRADES [74]	92.23%	66.58%
TRADES XCiT-L12 [33, 6]	91.73%	57.58%
Unlabeled data+TRADES [40]	91.10%	65.88%
TRADES [40]	85.29%	57.20%

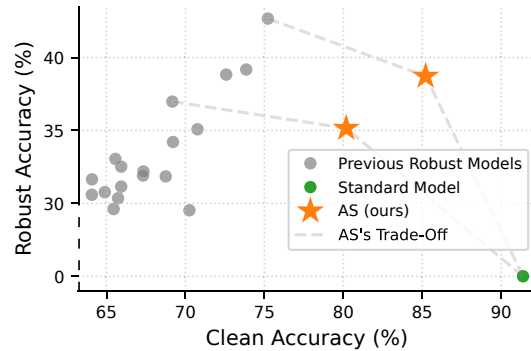


*: Uses “EDM + TRADES” [87] as the robust base model $h(\cdot)$.

†: AutoAttack raises the “potentially unreliable” flag (explained at the end of subsection 5.2.2), and adaptive attack reduces the attacked accuracy to 64.20%. AutoAttack does not raise this flag for our models.

CIFAR-100

Method	Clean	AA
AS (adaptive smoothing, ours) *	85.21%	38.72%
AS (adaptive smoothing, ours) **	80.18%	35.15%
Diffusion (EDM)+TRADES [87]	75.22%	42.67%
Unlabeled data+TRADES [40]	69.17%	36.98%
TRADES XCiT-L12 [33, 6]	70.76%	35.08%
Diffusion (DDPM)+TRADES [74]	63.56%	34.64%
SCORE Loss AT [69]	65.56%	33.05%
Diffusion (DDPM)+AT [78]	65.93%	31.15%
TRADES [40]	60.86%	30.03%



*: Uses “EDM+TRADES” [87] as the robust base model $h(\cdot)$.

**:: Uses “Unlabeled data+TRADES” [40] as the robust base model $h(\cdot)$.

While the above results demonstrate reconciled accuracy and robustness, the clean accuracy improvement over existing works may not seem highly prominent. Note that our method is still highly effective in this setting, but its efficacy is not fully reflected in the numbers. This is because SOTA robust base classifiers are already highly accurate on the easier CIFAR-10 dataset, almost matching standard models’ clean accuracy [73, 38, 39], leaving not much room for improvements. However, the accuracy-robustness trade-off remains highly penalizing for more challenging tasks such as CIFAR-100, for which existing robust models suffer significant accuracy degradation. As existing methods for improving standard model accuracy may not readily extend to robust ones, training-based trade-off alleviation struggles on harder tasks, making it particularly advantageous to mix already-trained classifiers via adaptive smoothing. We now support this claim with more significant improvements on CIFAR-100.

For CIFAR-100, we consider two robust base models and build two adaptive smoothing mixed classifiers. Compared with their corresponding robust base models, both mixed classifiers improve the clean accuracy by ten percentage points while only losing four points in

AutoAttacked accuracy. As of the submission of this paper, the mixed classifier whose robust base model is from [86] achieved an AutoAttacked accuracy better than any other methods on RobustBench [29], except [86] itself. Simultaneously, this mixed model offers a clean accuracy improvement of ten percentage points over any other listed models. These results confirm that adaptive smoothing significantly alleviates the accuracy-robustness trade-off.

We also report that the SA component of AutoAttack, which performs gradient-free black-box attacks on images that gradient-based attack methods fail to perturb, only changes very few predictions. Specifically, AutoAttack will raise a “potentially unreliable” flag if SA further reduces the accuracy by at least 0.2 percentage points. This flag is not thrown for our models in Table 4, indicating that the mixed classifiers’ robustness is not a result of gradient obfuscation. Thus, gradient-based attacks in AutoAttack sufficiently evaluate our models.

6. Conclusions. This paper proposes “adaptive smoothing”, a flexible framework that leverages the mixture of the output probabilities from an accurate model and a robust model to mitigate the accuracy-robustness trade-off of neural classifiers. We use theoretical and empirical observations to motivate our design, and mathematically prove that the resulting mixed classifier can inherit the robustness of the robust base model under realistic assumptions. We then adapt an adversarial input detector into a (deterministic) mixing network, further improving the accuracy-robustness trade-off. Solid empirical results confirm that our method can simultaneously benefit from the high accuracy of modern pretrained standard (nonrobust) models and the robustness achieved via SOTA robust classification methods.

Because our theoretical studies demonstrate the feasibility of leveraging the mixing network to eliminate the accuracy-robustness trade-off, future advancements in adversary detection can further reconcile this trade-off via our framework. Moreover, the proposed method conveniently extends to various robust base models and attack types/budgets. Thus, this work paves the way for future research to focus on accuracy or robustness without sacrificing the other.

REFERENCES

- [1] G. ADAM AND R. SPECIEL, *Evaluating Ensemble Robustness against Adversarial Attacks*, preprint, [arXiv:2005.05750](https://arxiv.org/abs/2005.05750), 2020.
- [2] M. A. AHMADI, R. DIANAT, AND H. AMIRKHANI, *An adversarial attack detection method in deep neural networks based on re-attacking approach*, *Multimedia Tools Appl.*, 80 (2021), pp. 10985–11014.
- [3] M. ALAM, S. DATTA, D. MUKHOPADHYAY, A. MONDAL, AND P. P. CHAKRABARTI, *Resisting Adversarial Attacks in Deep Neural Networks Using Diverse Decision Boundaries*, preprint, [arXiv:2208.08697](https://arxiv.org/abs/2208.08697), 2022.
- [4] A. ALDAHDOOH, W. HAMIDOUCE, AND O. DÉFORGES, *Revisiting model’s uncertainty and confidences for adversarial example detection*, *Appl. Intell.*, 53 (2023), pp. 509–531.
- [5] A. ALDAHDOOH, W. HAMIDOUCE, S. A. FEZZA, AND O. DÉFORGES, *Adversarial example detection for DNN models: A review and experimental comparison*, *Artif. Intell. Rev.*, 55 (2022), pp. 4403–4462.
- [6] A. ALI, H. TOUVRON, M. CARON, P. BOJANOWSKI, M. DOUZE, A. JOULIN, I. LAPTEV, N. NEVEROVA, G. SYNNAEVE, J. VERBEEK, AND H. JÉGOU, *XCiT: Cross-covariance image transformers*, in *Advances in Neural Information Processing Systems* 34, 2021, pp. 20014–20027.
- [7] B. G. ANDERSON, Z. MA, J. LI, AND S. SOJOUDI, *Tightened convex relaxations for neural network robustness certification*, in *2020 59th IEEE Conference on Decision and Control*, 2020, pp. 2190–2197.

- [8] B. G. ANDERSON AND S. SOJOUDI, *Certified robustness via locally biased randomized smoothing*, in 4th Annual Learning for Dynamics and Control Conference, PMLR 168, 2022, pp. 207–220.
- [9] B. G. ANDERSON AND S. SOJOUDI, *Data-driven certification of neural networks with random input noise*, *Trans. Control Netw. Syst.*, 10 (2023), pp. 249–260.
- [10] M. ANDRIUSHCHENKO, F. CROCE, N. FLAMMARION, AND M. HEIN, *Square Attack: A query-efficient black-box adversarial attack via random search*, in European Conference on Computer Vision, 2020, pp. 484–501.
- [11] A. ATHALYE, N. CARLINI, AND D. WAGNER, *Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples*, in 35th International Conference on Machine Learning, 2018, pp. 274–283.
- [12] A. ATHALYE, L. ENGSTROM, A. ILYAS, AND K. KWOK, *Synthesizing robust adversarial examples*, in 35th International Conference on Machine Learning, 2018.
- [13] T. BAI, J. LUO, J. ZHAO, B. WEN, AND Q. WANG, *Recent advances in adversarial training for adversarial robustness*, in Thirtieth International Joint Conference on Artificial Intelligence, 2021, pp. 4312–4321.
- [14] Y. BAI, B. G. ANDERSON, AND S. SOJOUDI, *Mixing classifiers to alleviate the accuracy-robustness trade-off*, in 6th Annual Learning for Dynamics & Control Conference (L4DC 2024), to appear.
- [15] Y. BAI, U. GARG, A. SHANKER, H. ZHANG, S. PARAJULI, E. BAS, I. FILIPOVIC, A. N. CHU, E. D. FOMITCHEVA, E. BRANSON, A. KIM, S. SOJOUDI, AND K. CHO, *Let's Go Shopping (LGS) – Web-Scale Image-Text Dataset for Visual Concept Understanding*, preprint, [arXiv:2401.04575](https://arxiv.org/abs/2401.04575), 2024.
- [16] Y. BAI, T. GAUTAM, Y. GAI, AND S. SOJOUDI, *Practical convex formulation of robust one-hidden-layer neural network training*, in American Control Conference, 2022, pp. 1535–1542.
- [17] Y. BAI, T. GAUTAM, AND S. SOJOUDI, *Efficient global optimization of two-layer ReLU networks: Quadratic-time algorithms and adversarial training*, *SIAM J. Math. Data Sci.*, 5 (2023), pp. 446–474, <https://doi.org/10.1137/21M1467134>.
- [18] Y. BALAJI, T. GOLDSTEIN, AND J. HOFFMAN, *Instance Adaptive Adversarial Training: Improved Accuracy Tradeoffs in Neural Nets*, preprint, [arXiv:1910.08051](https://arxiv.org/abs/1910.08051), 2019.
- [19] J. BORKAR AND P.-Y. CHEN, *Simple Transparent Adversarial Examples*, preprint, [arXiv:2105.09685](https://arxiv.org/abs/2105.09685), 2021.
- [20] N. CARLINI, F. TRAMER, K. D. DVIJOTHAM, L. RICE, M. SUN, AND J. Z. KOLTER, *(Certified!!) Adversarial robustness for free!*, in International Conference on Learning Representations, 2022.
- [21] N. CARLINI AND D. WAGNER, *Adversarial examples are not easily detected: Bypassing ten detection methods*, in ACM Workshop on Artificial Intelligence and Security, 2017, pp. 3–14.
- [22] N. CARLINI AND D. A. WAGNER, *Towards evaluating the robustness of neural networks*, in IEEE Symposium on Security and Privacy, 2017, pp. 39–57.
- [23] F. CARRARA, F. FALCHI, R. CALDELLI, G. AMATO, AND R. BECARELLI, *Adversarial image detection in deep neural networks*, *Multimedia Tools Appl.*, 78 (2019), pp. 2815–2835.
- [24] T. CHEN, S. LIU, S. CHANG, Y. CHENG, L. AMINI, AND Z. WANG, *Adversarial robustness: From self-supervised pre-training to fine-tuning*, in Conference on Computer Vision and Pattern Recognition, 2020, pp. 699–708.
- [25] T. CHEN, Z. ZHANG, S. LIU, S. CHANG, AND Z. WANG, *Robust overfitting may be mitigated by properly learned smoothing*, in International Conference on Learning Representations, 2020.
- [26] M. CHENG, Q. LEI, P. Y. CHEN, I. DHILLON, AND C. J. HSIEH, *CAT: Customized adversarial training for improved robustness*, in International Joint Conference on Artificial Intelligence, 2022, pp. 673–679.
- [27] K. T. CO, D. MARTINEZ-REGO, Z. HAU, AND E. C. LUPU, *Jacobian ensembles improve robustness trade-offs to adversarial attacks*, in Artificial Neural Networks and Machine Learning – ICANN 2022, Springer, 2022, pp. 680–691.
- [28] J. COHEN, E. ROSENFELD, AND Z. KOLTER, *Certified adversarial robustness via randomized smoothing*, in International Conference on Machine Learning, 2019, pp. 1310–1320.
- [29] F. CROCE, M. ANDRIUSHCHENKO, V. SEHWAG, E. DEBENEDETTI, N. FLAMMARION, M. CHIANG, P. MITTAL, AND M. HEIN, *Robustbench: A standardized adversarial robustness benchmark*, in Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.
- [30] F. CROCE, S. GOWAL, T. BRUNNER, E. SHEHAMER, M. HEIN, AND T. CEMGIL, *Evaluating the adversarial robustness of adaptive test-time defenses*, in International Conference on Machine Learning, 2022, pp. 4421–4435.

- [31] F. CROCE AND M. HEIN, *Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks*, in International Conference on Machine Learning, 2020, pp. 2196–2205.
- [32] F. CROCE AND M. HEIN, *Minimally distorted adversarial examples with a fast adaptive boundary attack*, in International Conference on Machine Learning, 2020, pp. 2206–2216.
- [33] E. DEBENEDETTI, V. SEHWAG, AND P. MITTAL, *A light recipe to train robust vision transformers*, in Conference on Secure and Trustworthy Machine Learning, 2023, pp. 225–253.
- [34] L. FAN, S. LIU, P.-Y. CHEN, G. ZHANG, AND C. GAN, *When does contrastive learning preserve adversarial robustness from pretraining to finetuning?*, in Advances in Neural Information Processing Systems 34, 2021, pp. 21480–21492.
- [35] M. FAZLYAB, A. ROBEY, H. HASSANI, M. MORARI, AND G. PAPPAS, *Efficient and accurate estimation of Lipschitz constants for deep neural networks*, in Annual Conference on Neural Information Processing Systems 32, 2019, pp. 11423–11434.
- [36] I. FURSOV, A. ZAYTSEV, P. BURNYSHEV, E. DMITRIEVA, N. KLYUCHNIKOV, A. KRAVCHENKO, E. ARTEMOVA, E. KOMLEVA, AND E. BURNAEV, *A differentiable language model adversarial attack on text classifiers*, IEEE Access, 10 (2022), pp. 17966–17976.
- [37] I. J. GOODFELLOW, J. SHLENS, AND C. SZEGEDY, *Explaining and harnessing adversarial examples*, in International Conference on Learning Representations, 2015.
- [38] S. GOWAL, C. QIN, J. UESATO, T. MANN, AND P. KOHLI, *Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples*, preprint, [arXiv:2010.03593](https://arxiv.org/abs/2010.03593), 2020.
- [39] S. GOWAL, S.-A. REBUFFI, O. WILES, F. STIMBERG, D. A. CALIAN, AND T. MANN, *Improving robustness using generated data*, in Annual Conference on Neural Information Processing Systems, 2021, pp. 4218–4233.
- [40] S. GOWAL, J. UESATO, C. QIN, P.-S. HUANG, T. MANN, AND P. KOHLI, *An Alternative Surrogate Loss for PGD-Based Adversarial Testing*, preprint, [arXiv:1910.09338](https://arxiv.org/abs/1910.09338), 2019.
- [41] K. HE, X. CHEN, S. XIE, Y. LI, P. DOLLÁR, AND R. GIRSHICK, *Masked autoencoders are scalable vision learners*, in Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
- [42] K. HE, X. ZHANG, S. REN, AND J. SUN, *Deep residual learning for image recognition*, in Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [43] M. HEIN AND M. ANDRIUSHCHENKO, *Formal guarantees on the robustness of a classifier against adversarial manipulation*, in Advances in Neural Information Processing Systems 30, 2017, pp. 2266–2276.
- [44] T.-K. HU, T. CHEN, H. WANG, AND Z. WANG, *Triple wins: Boosting accuracy, robustness and efficiency together by enabling input-adaptive inference*, in International Conference on Learning Representations, 2020.
- [45] S. H. HUANG, N. PAPERNOT, I. J. GOODFELLOW, Y. DUAN, AND P. ABBEEL, *Adversarial attacks on neural network policies*, in International Conference on Learning Representations, 2017.
- [46] A. ILYAS, L. ENGSTROM, A. ATHALYE, AND J. LIN, *Black-box adversarial attacks with limited queries and information*, in International Conference on Machine Learning, 2018, pp. 2137–2146.
- [47] X. JIA, Y. ZHANG, B. WU, K. MA, J. WANG, AND X. CAO, *LAS-AT: Adversarial training with learnable attack strategy*, in Conference on Computer Vision and Pattern Recognition, 2022, pp. 13398–13408.
- [48] M. JORDAN AND A. G. DIMAKIS, *Exactly computing the local Lipschitz constant of ReLU networks*, in Advances in Neural Information Processing Systems 33, 2020, pp. 7344–7353.
- [49] Q. KANG, Y. SONG, Q. DING, AND W. P. TAY, *Stable neural ODE with Lyapunov-stable equilibrium points for defending against adversarial attacks*, in Advances in Neural Information Processing Systems 34, 2021, pp. 14925–14937.
- [50] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, in International Conference on Learning Representations, 2015.
- [51] A. KOLESNIKOV, L. BEYER, X. ZHAI, J. PUIGSERVER, J. YUNG, S. GELLY, AND N. HOULSBY, *Big transfer (BiT): General visual representation learning*, in European Conference on Computer Vision, 2020, pp. 491–507.
- [52] A. KRIZHEVSKY, *Learning Multiple Layers of Features from Tiny Images*, Technical report, Department of Computer Science, University of Toronto, 2009, <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- [53] A. KUMAR, T. MA, P. LIANG, AND A. RAGHUNATHAN, *Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift*, in Conference on Uncertainty in Artificial Intelligence, 2022, pp. 1041–1051.

- [54] A. KURAKIN, I. J. GOODFELLOW, AND S. BENGIO, *Adversarial machine learning at scale*, in International Conference on Learning Representations, 2017.
- [55] A. LAMB, V. VERMA, J. KANNALA, AND Y. BENGIO, *Interpolated Adversarial Training: Achieving robust neural networks without sacrificing too much accuracy*, in ACM Workshop on Artificial Intelligence and Security, 2019, pp. 95–103.
- [56] B. LI, C. CHEN, W. WANG, AND L. CARIN, *Certified adversarial robustness with additive noise*, in Advances in Neural Information Processing Systems 32, 2019, pp. 9464–9474.
- [57] L. LI, Y. WANG, C. SITAWARIN, AND M. SPRATLING, *OODRobustBench: Benchmarking and Analyzing Adversarial Robustness under Distribution Shift*, preprint, [arXiv:2310.12793](https://arxiv.org/abs/2310.12793), 2023.
- [58] C. LIU, Y. DONG, W. XIANG, X. YANG, H. SU, J. ZHU, Y. CHEN, Y. HE, H. XUE, AND S. ZHENG, *A Comprehensive Study on Robustness of Image Classification Models: Benchmarking and Rethinking*, preprint, [arXiv:2302.14301](https://arxiv.org/abs/2302.14301), 2023.
- [59] X. LIU, M. CHENG, H. ZHANG, AND C.-J. HSIEH, *Towards robust neural networks via random self-ensemble*, in European Conference on Computer Vision, 2018, pp. 369–385.
- [60] Z. MA AND S. SOJOUDI, *A sequential framework towards an exact SDP verification of neural networks*, in International Conference on Data Science and Advanced Analytics, 2021, pp. 1–8.
- [61] A. MADRY, A. MAKELOV, L. SCHMIDT, D. TSIPRAS, AND A. VLADU, *Towards deep learning models resistant to adversarial attacks*, in International Conference on Learning Representations, 2018.
- [62] J. H. METZEN, T. GENEWEIN, V. FISCHER, AND B. BISCHOFF, *On detecting adversarial perturbations*, in International Conference on Learning Representations, 2017.
- [63] S.-M. MOOSAVI-DEZFOOLI, A. FAWZI, AND P. FROSSARD, *DeepFool: A simple and accurate method to fool deep neural networks*, in Conference on Computer Vision and Pattern Recognition, 2016, pp. 2574–2582.
- [64] S.-M. MOOSAVI-DEZFOOLI, A. FAWZI, J. UESATO, AND P. FROSSARD, *Robustness via curvature regularization, and vice versa*, in Conference on Computer Vision and Pattern Recognition, 2019, pp. 9078–9086.
- [65] D. NA, *Pytorch Adversarial Training on CIFAR-10*, 2020, <https://github.com/ndb796/Pytorch-Adversarial-Training-CIFAR>.
- [66] M. PAGLIARDINI, G. MANUNZA, M. JAGGI, AND T. CHAVDAROVA, *Improved generalization-robustness trade-off via uncertainty targeted attacks*, in International Conference on Machine Learning, 2022.
- [67] T. PANG, M. LIN, X. YANG, J. ZHU, AND S. YAN, *Robustness and accuracy could be reconcilable by (proper) definition*, in International Conference on Machine Learning, 2022, pp. 17258–17277.
- [68] T. PANG, K. XU, C. DU, N. CHEN, AND J. ZHU, *Improving adversarial robustness via promoting ensemble diversity*, in International Conference on Machine Learning, 2019, pp. 4970–4979.
- [69] N. PAPERNOT, P. MCDANIEL, I. GOODFELLOW, S. JHA, Z. B. CELIK, AND A. SWAMI, *Practical black-box attacks against machine learning*, in ACM Asia Conference on Computer and Communications Security, 2017.
- [70] S. PFROMMER, B. ANDERSON, J. PIET, AND S. SOJOUDI, *Asymmetric certified robustness via feature-convex neural networks*, in Advances in Neural Information Processing Systems 36, 2023, pp. 52365–52400.
- [71] S. PFROMMER, B. G. ANDERSON, AND S. SOJOUDI, *Projected randomized smoothing for certified adversarial robustness*, Trans. Mach. Learn. Res., (2023), <https://openreview.net/pdf?id=FObkvLwNSo>.
- [72] A. RAGHUNATHAN, S. M. XIE, F. YANG, J. C. DUCHI, AND P. LIANG, *Understanding and mitigating the tradeoff between robustness and accuracy*, in International Conference on Machine Learning, 2020, pp. 7909–7919.
- [73] S.-A. REBUFFI, S. GOWAL, D. A. CALIAN, F. STIMBERG, O. WILES, AND T. MANN, *Fixing Data Augmentation to Improve Adversarial Robustness*, preprint, [arXiv:2103.01946](https://arxiv.org/abs/2103.01946), 2021.
- [74] S. SABOUR, Y. CAO, F. FAGHRI, AND D. J. FLEET, *Adversarial Manipulation of Deep Representations*, preprint, [arXiv:1511.05122](https://arxiv.org/abs/1511.05122), 2015.
- [75] H. SALMAN, J. LI, I. RAZENSHTYRN, P. ZHANG, H. ZHANG, S. BUBECK, AND G. YANG, *Provably robust deep learning via adversarially trained smoothed classifiers*, in Advances in Neural Information Processing Systems 32, 2019.
- [76] L. SCHMIDT, S. SANTURKAR, D. TSIPRAS, K. TALWAR, AND A. MADRY, *Adversarially robust generalization requires more data*, in Advances in Neural Information Processing Systems 31, 2018.

- [77] V. SEHWAG, S. MAHLOUJIFAR, T. HANDINA, S. DAI, C. XIANG, M. CHIANG, AND P. MITTAL, *Robust learning meets generative models: Can proxy distributions improve adversarial robustness?*, in International Conference on Learning Representations, 2022.
- [78] A. SHAFABI, M. NAJIBI, M. A. GHIASI, Z. XU, J. DICKERSON, C. STUDER, L. S. DAVIS, G. TAYLOR, AND T. GOLDSTEIN, *Adversarial training for free!*, in Advances in Neural Information Processing Systems 32, 2019.
- [79] Z. SHI, Y. WANG, H. ZHANG, J. Z. KOLTER, AND C.-J. HSIEH, *Efficiently computing local Lipschitz constants of neural networks via bound propagation*, in Advances in Neural Information Processing Systems 35, 2022, pp. 2350–2364.
- [80] D. TERJEK, *Adversarial Lipschitz regularization*, in International Conference on Learning Representations, 2020.
- [81] F. TRAMÈR, N. CARLINI, W. BRENDL, AND A. MADRY, *On adaptive attacks to adversarial example defenses*, in Advances in Neural Information Processing Systems 33, 2020, pp. 1633–1645.
- [82] F. TRAMÈR, A. KURAKIN, N. PAPERNOT, I. J. GOODFELLOW, D. BONEH, AND P. D. MCDANIEL, *Ensemble adversarial training: Attacks and defenses*, in International Conference on Learning Representations, 2018.
- [83] D. TSIPRAS, S. SANTURKAR, L. ENGSTROM, A. TURNER, AND A. MADRY, *Robustness may be at odds with accuracy*, in International Conference on Learning Representations, 2019.
- [84] H. WANG, T. CHEN, S. GUI, T. HU, J. LIU, AND Z. WANG, *Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free*, in Advances in Neural Information Processing Systems 33, 2020, pp. 7449–7461.
- [85] J. WANG AND H. ZHANG, *Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks*, in International Conference on Computer Vision, 2019, pp. 6629–6638.
- [86] Z. WANG, T. PANG, C. DU, M. LIN, W. LIU, AND S. YAN, *Better diffusion models further improve adversarial training*, in International Conference on Machine Learning, 2023, pp. 36246–36263.
- [87] Y.-Y. YANG, C. RASHTCHIAN, H. ZHANG, R. R. SALAKHUTDINOV, AND K. CHAUDHURI, *A closer look at accuracy vs. robustness*, in Advances in Neural Information Processing Systems 33, 2020, pp. 8588–8601.
- [88] Y. YU, X. GAO, AND C.-Z. XU, *MORA: Improving ensemble robustness evaluation with model-reweighing attack*, in Advances in Neural Information Processing Systems 35, 2022, pp. 26955–26965.
- [89] H. ZHANG AND J. WANG, *Defense against adversarial attacks using feature scattering-based adversarial training*, in Advances in Neural Information Processing Systems 32, 2019.
- [90] H. ZHANG, Y. YU, J. JIAO, E. P. XING, L. E. GHAOUI, AND M. I. JORDAN, *Theoretically principled trade-off between robustness and accuracy*, in International Conference on Machine Learning, 2019, pp. 7472–7482.
- [91] S. ZHAO, X. WANG, AND X. WEI, *Mitigating the Accuracy-Robustness Trade-Off via Multi-teacher Adversarial Distillation*, preprint, [arXiv:2306.16170](https://arxiv.org/abs/2306.16170), 2023.
- [92] H. ZHENG, Z. ZHANG, J. GU, H. LEE, AND A. PRAKASH, *Efficient adversarial training with transferable adversarial examples*, in Conference on Computer Vision and Pattern Recognition, 2020.
- [93] Y. ZHENG, R. ZHANG, AND Y. MAO, *Regularizing neural networks via adversarial model perturbation*, in Conference on Computer Vision and Pattern Recognition, 2021, pp. 8156–8165.