Focus on Hiders: Exploring Hidden Threats for Enhancing Adversarial Training

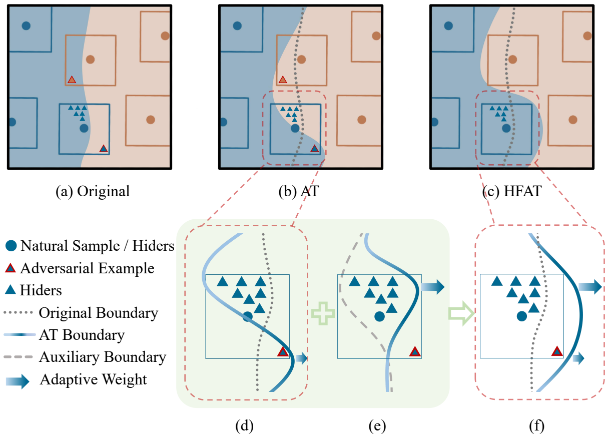
Qian Li[1,2], Yuxiao Hu[2,3], Yinpeng Dong[4], Dongxiao Zhang[2], Yuntian Chen[2*]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3]The Hong Kong Polytechnic University, HongKong, China
[4]Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua-Bosch Joint ML Center, THBI Lab, BNRist Center, Tsinghua University, Beijing 100084, China
qianl01205@sjtu.edu.cn, huyuxiao20@mails.ucas.ac.cn
dongyinpeng@mail.tsinghua.edu.cn, {dzhang, ychen}@eitech.edu.cn

Abstract

Adversarial training is often formulated as a min-max problem, however, concentrating only on the worst adversarial examples causes alternating repetitive confusion of the model, i.e., previously defended or correctly classified samples are not defensible or accurately classifiable in subsequent adversarial training. We characterize such non-ignorable samples as “hiders”, which reveal the hidden high-risk regions within the secure area obtained through adversarial training and prevent the model from finding the real worst cases. We demand the model to prevent hiders when defending against adversarial examples for improving accuracy and robustness simultaneously. By rethinking and redefining the min-max optimization problem for adversarial training, we propose a generalized adversarial training algorithm called ***H****ider-****F****ocused* ***A****dversarial* ***T****raining (HFAT). HFAT introduces the iterative evolution optimization strategy to simplify the optimization problem and employs an auxiliary model to reveal hiders, effectively combining the optimization directions of standard adversarial training and prevention hiders. Furthermore, we introduce an adaptive weighting mechanism that facilitates the model in adaptively adjusting its focus between adversarial examples and hiders during different training periods. We demonstrate the effectiveness of our method based on extensive experiments, and ensure that HFAT can provide higher robustness and accuracy.*

1. Introduction

Although deep neural networks (DNNs) have made significant progress in recent years [6, 11, 28], they are easily fooled by adversarial examples to make incorrect predictions [7, 9, 16, 18]. These malicious attacks pose a threat to the security and well-being of individuals, which highlights the importance of adversarial defense efforts. Among them, adversarial training is proven to be the most effective defense method against adversarial attacks [14, 18].

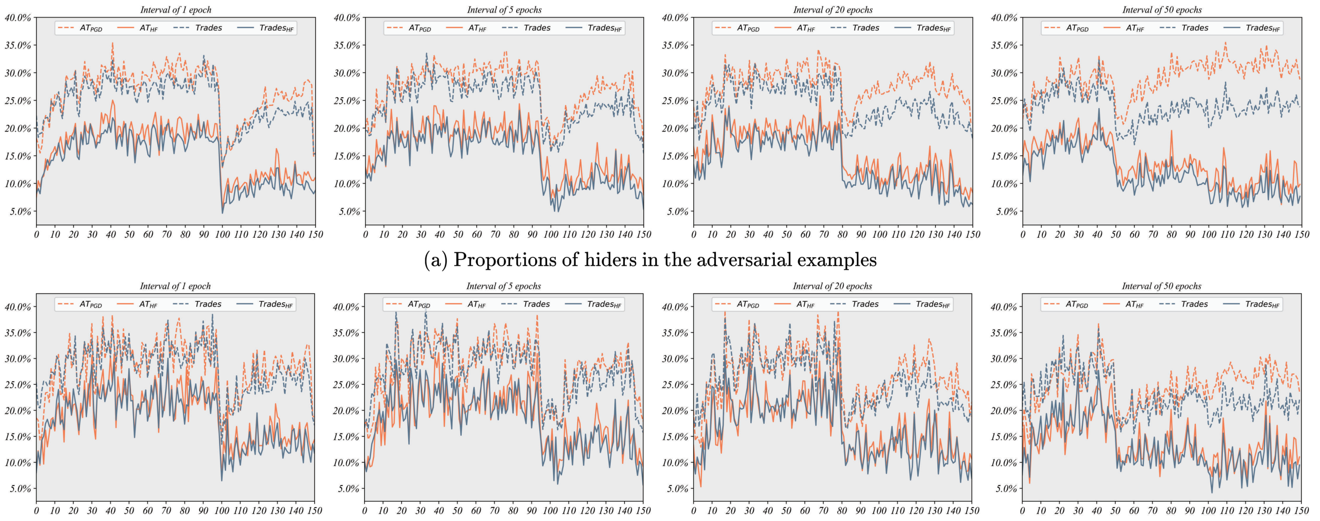


(a) Original (b) AT (c) HFAT

(d) (e) (f)

Figure 1. Illustration of our core idea. Previous adversarial training methods have single-mindedly concentrated on worst-case adversarial examples, aiming to accurate classification of such examples. However, these methods fail to protect hidden high-risk regions. We refer to these regions' samples as hiders (blue-filled triangles) which are correctly classified in the original model but misclassified after adversarial training due to excessive accommodation of adversarial examples (blue-filled triangles within *original boundary* and outside of *AT boundary*, as depicted in (b), (d)). It is noteworthy to mention that this phenomenon of diminished accuracy also affects natural samples (blue-filled circle), which can be considered as a special type of hiders. By introducing an auxiliary model that exposes the hidden high-risk regions where hiders are located (blue-filled triangles outside of *auxiliary boundary*, as depicted in (e)), we can obtain the optimization direction to prevent hiders. Our core idea is to adaptively defend against both adversarial examples and hiders simultaneously, which promises a defense mechanism that ensures superior robustness and accuracy.

[*] Corresponding author



(a) Proportions of hiders in the adversarial examples

(b) Proportions of hiders in the original examples

Figure 2. Visualization of the proportions of hiders across different epochs. The first row (a) denotes the ratios of hiders in the adversarial examples. The horizontal axis denotes the present epoch. Adversarial examples that are initially defended successfully in the present epoch, but thereafter fail after intervals of 1, 5, 20, and 50 epoch(s), are referred to as hiders. The variations in the proportion of these hiders are depicted in the graphs. We plot the proportions for four different methods: AT_{PGD}, AT_{HF}, Trades and $Trades_{HF}$ (subscript "HF" represents the proposed approach); The second row (b) follows the same process to depict proportions of hiders in the original samples. For original samples, hiders refer to samples that are first classified correctly in the present epoch but then fail after periods of 1, 5, 20, and 50 epochs. Full details are in Sec. 4.3.1.

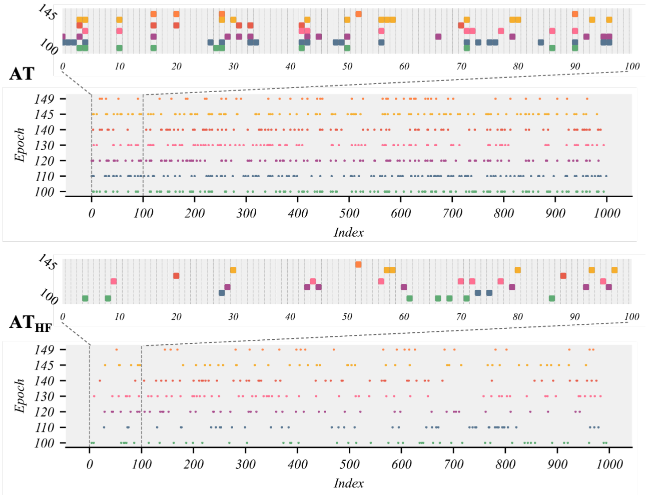


Figure 3. Statistical graph of hiders' occurrence locations. A set of 1000 adversarial examples, which the model fails to defend against at epoch 150 using PGD_{20}, is collected. If the models from prior epochs (100, 110, 120, 130, 140, 145, and 149) successfully defend these samples, they are labeled as hiders. Additionally, markers are placed at the respective indices. To enhance visualization, we magnify the indices in the range of 0-100 to better discern the disparities between AT_{PGD} and AT_{HF} (AT with our proposed method). Full details are in Sec. 4.3.1.

Following the min-max optimization problem [18] of adversarial training, previous methods [8, 18, 21, 29, 30, 34] always single-mindedly focus on the worst-case adversarial examples to obtain the optimal solution of inner problem. However, these methods tend to overlook the potential vulnerabilities that also exist in secure areas and result in compromised robustness and accuracy. Concretely, concentrating only on the worst-case adversarial examples causes alternating repetitive confusion of the model as seen in Fig. 1, *i.e.*, adversarial or natural samples that were defended or accurately categorized against in the preceding adversarial training epoch are no longer amenable to defense or accurate classification in the subsequent epoch. Fig. 2 shows the ubiquity of this phenomenon throughout all training epochs. Moreover, as the epoch interval increases, a greater number of previously defended samples or accurately classified samples will become susceptible to attacks or misclassification in subsequent epochs. Besides, the observed phenomenon exhibits intermittency. As depicted in Fig. 3, the decision boundary repeatedly confuses certain samples, thereby hindering the model from identifying the genuine worst case (the worst-case adversarial example we find in this epoch may be just a perturbed sample without attack performance for the model in the previous epoch) and impacts the model's performance. We analyze that this phenomenon is caused by single-minded optimization, which excessively adjusts the decision boundary towards adversarial examples and neglects focused protections for the temporary secure regions. This overlooked issue motivates us to reconsider the min-max optimization problem for adversarial training, and explore the possibilities of preventing potential dangers.

In this paper, we first define those non-negligible samples as "hiders", which were successfully defended or correctly classified in the previous epoch of adversarial training, but exhibit strong attack capability or are misclassified in the later epoch. We propose a new generalized adversarial training algorithm dubbed **H**iders-**F**ocused **A**dversarial **T**rainning (HFAT), as shown in Fig. 1 (c) and (f). HFAT enhances robustness and accuracy by preventing potentially vulnerable areas where hiders are most likely to be identified, while maintaining defense against adversarial examples. Specifically, we redefine the min-max optimization problem and propose the *iterative evolution optimization strategy* to simplify the problem, which allows us to only consider hiders that are relevant to the next epoch. More specifically, by exploring the intrinsic relationship between hiders and adversarial examples, we model the distribution of hiders as a priori knowledge. Given that hiders pose no immediate threat to the current model, directly utilizing them into adversarial training yields limited advantages. Therefore, through probabilistic sampling guided by the prior distribution, we train an auxiliary model that reveals hiders to determine the optimal direction for preventing hiders by adversarial training on auxiliary model. HFAT integrates the adversarial training optimization directions from both the standard model and the auxiliary model to jointly optimize the network. Besides, in order to better couple the two optimization directions during various training phases, we further design an adaptive weighting mechanism that adjusts the emphasis between hiders and adversarial examples in a dynamic manner. In short, the model can boost the optimization weight for an aspect depending on which aspect is currently more needed.

Our contributions are summarized as follows.

- We first reveal that the single-minded focus of adversarial training on adversarial examples neglects the hidden threats in secure regions that have been successfully defended in the current epoch, resulting in compromised robustness and accuracy.
- We define hiders and redefine the min-max optimization objective aiming to achieve better robustness and accuracy by preventing potentially vulnerable regions while defending against adversarial examples.
- We propose a generalized adversarial training strategy called Hiders-Focused Adversarial Training (HFAT). HFAT introduces the *iterative evolution optimization strategy* to simplify the optimization problem and employs an auxiliary model to reveal hiders, effectively combining the optimization directions of standard adversarial training and prevention of hiders. Besides, HFAT includes an adaptive weighting mechanism to improve the coupling of the two optimization objectives.
- We demonstrate the effectiveness of our method based on extensive experiments, and reveal that HFAT effectively mitigates hidden threats posed by hiders.

2. Background and Related work

2.1. Adversarial Attack

The objective of adversarial attacks is to exploit the model's vulnerability in the vicinity of decision boundaries by introducing small, imperceptible perturbations to the inputs, tricking the model into providing incorrect classifications or predictions. Adversarial attack can be represented as the following optimization objective:

$$\max_{\boldsymbol{\delta}\in\mathcal{B}(\epsilon)} \mathcal{L}_{\mathrm{CE}}(f_\theta(\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}),$$

where $\mathcal{L}_{\mathrm{CE}}$ denotes the cross-entropy loss function and $\mathcal{B}(\epsilon)=\{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|\leq\epsilon\}$ limits the perturbation $\boldsymbol{\delta}$ under a certain distance metric (usually ℓ_p-norm).

Several adversarial attack methodologies have been proposed, exposing the susceptible components of deep learning models. To increase the efficacy of adversarial perturbations, Projected Gradient Descent (PGD [18]) refines them iteratively. Carlini and Wagner's attack (C&W [3]) formulates an optimization problem with the goal of obtaining misclassification with the fewest possible perturbations. Momentum Iterative Method (MIM [7]) enhances traditional iterative optimization techniques by incorporating a momentum term, facilitating more effective and efficient exploration of the adversarial perturbation space. AutoAttack [4] presents a suite of diverse attack methods to evaluate model robustness comprehensively. The success of these attack methods makes the adversarial defense a meaningful work for improving model robustness.

2.2. Adversarial Training

Adversarial training is an essential approach to enhance the robustness of deep learning models against adversarial attacks. It involves augmenting the training dataset with adversarial examples, forcing the model to learn and defend against adversarial threats. The foundation of adversarial training lies in a min-max optimization framework, which can be formalized as:

$$\min_{\theta}\max_{\boldsymbol{\delta}\in\mathcal{B}(\epsilon)} \mathcal{L}_{\mathrm{CE}}(f_\theta(\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}).$$

Many noteworthy methods have been proposed within the framework of adversarial training. Madry et al. introduced the foundational $\mathrm{AT}_{\mathrm{PGD}}$ [18] framework, focusing on improving robustness of the models. An early stopping variant of $\mathrm{AT}_{\mathrm{PGD}}$ [23], proposed by Rice et al., demonstrated notable improvements. Zhang et al. presented the TRADES [34] method, exploring a trade-off between standard accuracy and adversarial robustness. Wu et al. delved into the weight loss landscape, introducing Adversarial Weight Perturbation (AWP [30]) to effectively enhance model robustness. MART [29], introduced by Wang et al., improved the adversarial example generation process by simultaneously incorporating misclassified clean examples. Jia et al. introduced Learnable Attack Policy Adversarial Training (LAS-AT [13]), a concept that involved learning to automatically generate better attack policies to enhance model robustness.

However, these methodologies still excessively emphasize adversarial examples [2, 22], neglecting the potential threats concealed within secure regions. As a result, these models are repeatedly confused by hiders and unable to identify the genuine worst case, leading to limited robustness and accuracy. In this paper, we enhance the robustness and accuracy of the model by directing our focus towards hidden threats from a new perspective.

3. Methodology

3.1. Hiders

While adversarial examples directly expose vulnerabilities in the current trained model, hiders reveal hidden threats within the decision boundaries of the model, *i.e.*, samples that were correctly classified or defended in the previous epoch of adversarial training cannot be accurately classified or defended in subsequent epochs. Besides, as illustrated in the second row of Fig. 2, it is noteworthy to emphasize that certain natural samples exhibit hider-like characteristics. Thus, by implementing proactive defense mechanisms against hiders during adversarial training, the accuracy and robustness of the model can be improved simultaneously.

Definitions of hiders. We first define the hider $\hat{\boldsymbol{x}} = \boldsymbol{x} + \hat{\boldsymbol{\delta}}^j$ of sample $(\boldsymbol{x}, \boldsymbol{y})$ in the current i-th epoch with respect to the later j-th epoch as follows:

$$D(f_{\theta^i}(\hat{\boldsymbol{x}})) = y, D(f_{\theta^j}(\hat{\boldsymbol{x}})) \neq y, \quad i, j \in \{1, 2, ...\}, j > i,$$

where $\hat{\boldsymbol{\delta}}^j \in \mathcal{B}(\epsilon) \bigcap S_i$, S_i indicates the interior of the decision boundary in i-th epoch, and $\hat{\boldsymbol{x}}$ is defended or correctly classified at the i-th epoch and fails at the j-th epoch. D is the classification function that maps the probability distribution $f_{\theta^i}(\hat{\boldsymbol{x}})$ to the class y with the highest probability.

Similarly to the adversarial examples, for the model of the i-th epoch, there exists a worst-case hider $\hat{\boldsymbol{x}}^* = \boldsymbol{x} + \hat{\boldsymbol{\delta}}^*$, which can be expressed as

$$(j^*, \hat{\boldsymbol{\delta}}^*) = \underset{j, \hat{\boldsymbol{\delta}}^j}{\operatorname{argmax}} \mathcal{L}_{\mathrm{CE}}(f_{\theta^j}(\boldsymbol{x} + \hat{\boldsymbol{\delta}}^j), \boldsymbol{y}). \tag{1}$$

Unlike the worst-case adversarial example, which indicates the sample with the strongest attack performance under the current model, the worst-case hider indicates the sample within the current decision boundary that exhibits the highest upper bound on its attack performance during the future epochs. As an illustration, let $\hat{\boldsymbol{x}}^*$ represent the worst-case hider of $\boldsymbol{x}$ at the i-th epoch, indicating that $\hat{\boldsymbol{x}}^*$ lies within the decision boundary under the model f_{θ^i}, and the maximum loss value of $\hat{\boldsymbol{x}}^*$ in the future j^*-th epoch is larger than the maximum loss value of any other samples (also within the decision boundary under the model f_{θ^i}) in future epochs.

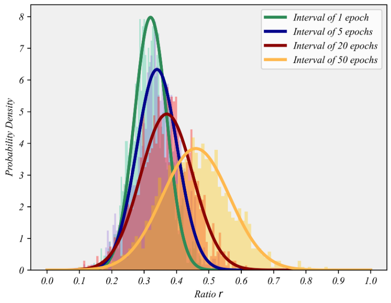


Figure 4. Histograms show the distribution of ratio values (r) from 10,000 samples across five defense models (AT$_{\mathrm{PGD}}$, Trades, MART, AWP, HELP), quantifying distances between hider and original samples versus adversarial and original samples, in the model's gradient direction. Colored curves represent Gaussian fits that correspond to the histograms at different interval epochs.

Empirical probability distribution of hiders. Due to the delayed threat of hiders, it is difficult to pinpoint worst-case hiders. Besides, the distribution of hiders depends on natural samples and models, which encourages us to explore the relative position of hiders from previous adversarial training models rather than absolute location information. We observed a remarkable similarity in the relative positional relationships between hiders and natural/adversarial examples across various adversarial training methods and training phases. This guides us to model the empirical probability distribution [19, 24, 26] of hiders' relative position. The distribution not only reveals the position of hiders with natural samples and adversarial examples, but also elucidates the probability that a sample belongs to hiders.

Based on the observations, we use the Gaussian distribution G to model the relative positional information. Specifically, at the i-th epoch, the positional information distribution of hiders relative to the j-th epoch is denoted as G_j. To obtain the regions where hiders are likely to occur, we compared the distances of both hiders and adversarial examples to the original samples, resulting in the relative position ratio r. Fig. 4 displays the histograms of the ratio r for 10000 hiders, which are computed on five defense models (AT$_{\mathrm{PGD}}$ [18], Trades [34], MART [29], AWP [30] and HELP [21]). We fit the data in accordance with the Gaussian distribution characteristics evident in the histogram. Besides, observation reveals that as the number of epochs increases, the mean and variance of the ratios r for the generated hiders also show an increase. Due to the statistical analysis being conducted on multiple models and a large number of sample points, the distribution characteristics of hiders that we observed exhibit universality. We can regard r as a hyperparameter obtained from priori knowledge that can provide assistance in defense against hiders.

3.2. Hider-Focused Adversarial Training (HFAT)

To enhance both the robustness and accuracy of the model, our objective is to proactively defend the worst cases involving adversarial examples and hiders. This dual focus can be expressed as the following optimization objective,

$$\begin{aligned}\min[&\max_{\boldsymbol{\delta}\in\mathcal{B}(\epsilon)} \mathcal{L}_{\mathrm{CE}}(f_{\theta^i}(\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y})+\\ &\max_{j,\hat{\boldsymbol{\delta}}^j\in S^i} \mathcal{L}_{\mathrm{CE}}(f_{\theta^j}(\boldsymbol{x}+\hat{\boldsymbol{\delta}}^j),\boldsymbol{y})],\end{aligned} \tag{2}$$

where the former is optimized for the adversarial examples, the latter for preventing the potential dangers of hiders, and j denotes that the worst-case hider reaches the maximum loss value at the j-th epoch.

The second term of optimization objective (2) forces us to find the maximum loss value of the samples $\boldsymbol{x}+\hat{\boldsymbol{\delta}}$ in all future epochs. We propose *Iterative Evolution Optimization Strategy* to simplify the problem. According to Theorem 1, we can optimize objective (2) by considering only the worst-case in the next epoch, *i.e*., optimization of objective (2) can be simplified into optimizing objective (3). The proof of Theorem 1 is available in the supporting material.

$$\begin{aligned}\min[&\max_{\boldsymbol{\delta}\in\mathcal{B}(\epsilon)} \mathcal{L}_{\mathrm{CE}}(f_{\theta^i}(\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y})+\\ &\max_{\hat{\boldsymbol{\delta}}^{i+1}\in S^i} \mathcal{L}_{\mathrm{CE}}(f_{\theta^{i+1}}(\boldsymbol{x}+\hat{\boldsymbol{\delta}}^{i+1}),\boldsymbol{y})],\end{aligned} \tag{3}$$

Theorem 1. *(Iterative Evolution Optimization Strategy) We can optimize objective (2) by iteratively optimizing against the worst-case hider for the next epoch.*

Auxiliary model. The second term of objective (3) serves the purpose of enabling the current model to defend against hidden threatening regions within the current decision boundary. However, optimizing the current model for future scenarios poses a challenge due to the inability to calculate the derivative of the objective (3)'s second term with respect to the current model's weight parameters θ^i. Moreover, if we directly approximate the second term with hiders' loss function $\mathcal{L}_{\mathrm{CE}}(f_{\theta^i}(\boldsymbol{x}+\hat{\boldsymbol{\delta}}^j))$ under model f_{θ^i}, the conventional gradient descent method encounters optimization difficulty since the hiders are not aggressive for model f_{θ^i}.

To address this issue, we obtain an auxiliary model that has a higher loss at $\boldsymbol{x}+\hat{\boldsymbol{\delta}}^j$, thereby exposing the region where hiders are located. Furthermore, we enhance the optimization of model f_{θ^i} by adding the optimization direction from the adversarial training on auxiliary model as momentum. This compels the current model to acquire optimization directions that effectively prevent hiders.

To make the auxiliary model $f_{\hat{\theta}^i}$ expose the region where hiders are situated, it is necessary to determine where hiders are most probable to emerge. By sampling from the empirical probability distribution G, we can determine the relative location ratio r of hiders between natural samples and adversarial examples. This allows us to find the most probable region where hiders are placed, as sampling is dependent on probability. In particular, since we only need to consider the hiders associated with the next epoch, we sample the relative position ratio r from G_1. Then we can obtain the auxiliary model $f_{\hat{\theta}^i}$ through reverse training as follows:

$$\hat{\theta}^i \leftarrow \theta^i + \eta\nabla_{\theta^i}\mathcal{L}_{\mathrm{CE}}(f_{\theta^i}(T(\boldsymbol{x},\boldsymbol{x}_{\mathrm{adv}},r)),\boldsymbol{y}), \quad r\sim G_1,$$

where η is the learning rate, T represents a position transformation function that computes the most probable regions for hiders based on sampled r, natural samples, and adversarial examples. We incorporate a minor amount of noise within ϵ into the process to introduce a degree of randomness. Furthermore, by applying adversarial training on auxiliary model, we are able to determine the gradient direction of defending the hidden threatening regions, which can be utilized as an approximation for the second term in objective (3). We introduce it as momentum p in the training of model f_{θ^i}, where the p can be denoted as:

$$\begin{aligned}p^i &= \nabla_{\hat{\theta}^i}(\mathcal{L}_{\mathrm{CE}}(f_{\hat{\theta}^i}(\boldsymbol{x}+\boldsymbol{\delta}^*),\boldsymbol{y})),\\ \boldsymbol{\delta}^* &= \max_{\boldsymbol{\delta}^*\in\mathcal{B}(\epsilon)} \mathcal{L}_{\mathrm{CE}}(f_{\hat{\theta}^i}(\boldsymbol{x}+\boldsymbol{\delta}^*),\boldsymbol{y}).\end{aligned} \tag{4}$$

HFAT aims to mitigate potential threats while defending adversarial examples, so the optimization can be expressed as a coupling of standard adversarial training and auxiliary model guided optimization, which can be formalized as:

$$\theta^{i+1} \leftarrow \theta^i - \eta(\nabla_{\theta^i}\mathcal{L}_{\mathrm{CE}}(f_{\theta^i}(\boldsymbol{x}+\boldsymbol{\delta}),y)+p^i). \tag{5}$$

Adaptive weighting mechanism. In fact, HFAT can be conceptualized as consisting of two adversarial training branches. The first branch involves standard adversarial training, where the model is trained to focus on the worst-case adversarial examples. The second branch concentrates on adversarial training the auxiliary model to assist the model in defending against the region with the highest hidden threat. Considering that the threat intensity of adversarial examples and hiders to the model varies across samples and training phases, we devise an adaptive weighting mechanism to improve the coupling between the two adversarial training branches.

We utilize the disparity between the outputs of natural and adversarial examples as a metric, which represents significance of the branch. If there is a significant disparity between the two outputs, it indicates that the branch is comparatively undertrained and requires increased emphasis during training. The Kullback-Leibler divergence is used to quantify the difference. The adaptive weighting mechanism can be expressed as:

$$\lambda_A = \frac{e^{\mathrm{KL}(f_{\hat{\theta}}(\boldsymbol{x})||f_{\hat{\theta}}(\boldsymbol{x}'))}}{e^{\mathrm{KL}(f_{\theta}(\boldsymbol{x})||f_{\theta}(\boldsymbol{x}'))}+e^{\mathrm{KL}(f_{\hat{\theta}}(\boldsymbol{x})||f_{\hat{\theta}}(\boldsymbol{x}'))}}, \tag{6}$$

where λ_A denotes the weight of the momentum p from the auxiliary model's branch, and the weight of standard adversarial training's branch is $\lambda_S = 1 - \lambda_A$.

Training strategy. With the introduction of auxiliary model and adaptive weighting mechanisms, the update of HFAT's weighting parameters finally can be represented as formula 7, where p^i is shown in formula 4.

$$\theta^{i+1} \leftarrow \theta^i - \eta(\lambda_S \nabla_{\theta^i} \mathcal{L}_{\mathrm{CE}}(f_{\theta^i}(\boldsymbol{x}+\boldsymbol{\delta}), y) + \lambda_A p^i). \quad (7)$$

4. Experiments

4.1. Experimental setting

Dataset: We conduct extensive experiments on the CIFAR-10 [15], CIFAR-100 [15], and SVHN [20] datasets. We employ a perturbation budget of $8/255$ for three datasets. **Network Architectures:** To train on these datasets, we employed a standard network Pre-ResNet18 [12] and an advanced large-scale network (WideResNet-34-10 [32]). **Baselines:** We adopt a standard defense baseline: $\mathrm{AT}_{\mathrm{PGD}}$ [18] and four strong defense baselines: TRADES [34], MART [29], AWP [30], HELP [21] to verify generalization in various conditions. **Training Details**: All defenses undergo 200 epochs of training using SGD with a momentum of 0.9, weight decay of 5×10^{-4}, and an initial learning rate of 0.1. The learning rate is reduced by a factor of 10 at the 100-th and 150-th epoch. Simple data augmentations, including a 32×32 random crop with 4-pixel padding and random horizontal flip, are applied during the training process for all methods.

4.2. Performance analysis

4.2.1 Performance on robustness and accuracy

We utilize two standard attacks, namely FGSM [9] and PGD [18], as well as four strong attacks: C&W [3], MIM [7], $\mathrm{AA}_{\mathrm{rand}}$ [4] (composed of APGD-CE [4] and APGD-DLR [4]) and $\mathrm{AA}_{\mathrm{standard}}$ [4] (a collection of diverse parameter-free attacks consisting three white-box attacks: APGD-CE [4] and APGD-T [4] and FAB-T [5], and a black-box attack: Square Attack [1]). C&W uses the margin-based loss function described in [3] and utilizes PGD for optimization. Specifically, we employ 20 and 100 steps for PGD, 20 steps for MIM, and 30 steps for C&W. The step size for these attacks is set to $\alpha = \varepsilon/4$. Tab. 1 showcases the performance enhancement achieved by HFAT across the five defense baselines on CIFAR-10 [15]. Results on the CIFAR-100 and SVHN datasets will be presented in the supporting material. We observe that HFAT improves almost both natural accuracy and robust accuracy against various attack methods, confirming the efficiency and applicability of our strategy.

4.2.2 Performance under black-box attacks

Despite the inclusion of black-box attacks in $\mathrm{AA}_{\mathrm{standard}}$, we further extend our evaluation to encompass transfer-based black-box attacks [25, 27, 31] using PGD-20. We employed adversarial examples generated from the source model to attack the target model. The results in Tab. 2 show that the model improved by HFAT achieved better transferability in adversarial attacks (evident in the red box region in the figure, where the HFAT model exhibited a higher success rate when attacking the same target model). Additionally, it demonstrated superior performance in defending against transfer adversarial examples (as depicted by the yellow background region in the figure, where the HFAT model exhibited higher accuracy when faced with adversarial examples from the same source model).

4.3. Analysis of hiders

4.3.1 Defense Performance

We illustrate HFAT's defensive performance against hiders through two aspects.

We first verify that HFAT can effectively defend against potential threats from hiders. We compute and plot the proportions of hiders at different intervals (interval 1, 5, 20, and 50 epochs) in Fig. 2. It is observed that our method significantly reduces the proportion of hiders compared to $\mathrm{AT}_{\mathrm{PGD}}$ and Trades defenses alone. Furthermore, as the interval value increases, the proportion of adversarial hiders for $\mathrm{AT}_{\mathrm{PGD}}$ and Trades increases, while our method shows almost no difference across different intervals. This indicates that training the model using experience distribution sampling with an interval of 1 epoch is reasonable and ultimately enables better generalization in defending against hiders with larger interval values.

We next verify that HFAT prevents repeated threats to the model by hiders. We visualize the specific index positions at which hiders appear in Fig. 3 and observe a phenomenon of repeated occurrences in the $\mathrm{AT}_{\mathrm{PGD}}$. This repetition appears to follow an intermittent and alternating pattern. However, in the $\mathrm{AT}_{\mathrm{HF}}$, hiders generally do not exhibit repeated occurrences, indicating that HFAT can effectively and persistently defend against hiders.

4.3.2 Loss landscape

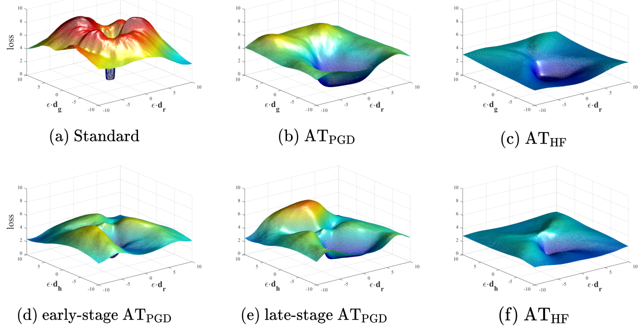
In this subsection, we directly validate the effectiveness of HFAT by visualizing the loss landscape as shown in Fig. 5 [8, 10, 17]. Firstly, we perturb the input in the gradient direction and random direction as illustrated in Fig. 5 (a)-(c) [8]. It's obvious that HFAT flattens the loss landscape more significantly compared to $\mathrm{AT}_{\mathrm{PGD}}$, which illustrates that HFAT provides better robustness. Additionally, we visualize the loss landscape along the hider's direction and random direction in Fig. 5 (d)-(f). It can be seen through Fig. 5 (e) that $\mathrm{AT}_{\mathrm{PGD}}$ exhibits local peaks, which confirms

Table 1. Comparison of performance improvement of HFAT applied to five different baselines on the CIFAR-10 dataset and implemented them on the PreAct ResNet-18 and WideResNet34-10 architectures. For testing the attack methods, we selected FGSM, PGD_{20}, PGD_{100}, CW, MIM, AA_{rand}, and AA. Here, AA refers to the standard version of AutoAttack.

	PreAct ResNet-18								WideResNet34-10							
	Natural	FGSM	PGD_{20}	PGD_{100}	CW	MIM	AA_{rand}	AA	Natural	FGSM	PGD_{20}	PGD_{100}	CW	MIM	AA_{rand}	AA
AT_{PGD}	81.66	57.51	52.64	52.51	50.29	52.83	49.04	48.25	86.40	61.83	55.60	55.19	54.59	55.70	52.95	52.06
AT_{HF}	**81.88**	**60.04**	**56.39**	**56.23**	**52.60**	**56.46**	**51.64**	**50.61**	**87.53**	**65.76**	**60.06**	**59.87**	**57.64**	**60.18**	**56.55**	**55.58**
TRADES	**81.24**	59.24	55.71	55.37	50.45	55.63	49.95	49.20	83.68	61.78	59.31	59.25	54.29	59.31	54.02	53.46
$\text{TRADES}_{\text{HF}}$	80.39	**59.61**	**57.41**	**57.12**	**51.20**	**56.95**	**50.96**	**50.35**	**85.38**	**63.80**	**61.12**	**61.03**	**55.88**	**61.16**	**55.62**	**55.05**
MART	80.63	59.54	56.16	55.86	50.31	55.41	50.47	49.62	83.98	61.32	58.43	58.12	54.74	58.07	53.64	52.36
MART_{HF}	**81.14**	**59.73**	**57.24**	**56.97**	**51.11**	**56.36**	**51.61**	**50.74**	**84.76**	**64.03**	**61.63**	**61.47**	**56.18**	**60.73**	**55.23**	**54.77**
AWP	80.81	59.38	55.59	55.47	51.89	55.70	51.04	50.06	85.65	62.75	58.82	58.69	55.56	59.24	55.39	53.61
AWP_{HF}	**81.17**	**59.83**	**55.95**	**55.87**	**52.31**	**56.27**	**51.82**	**50.28**	**86.41**	**64.18**	**62.23**	**62.06**	**57.42**	**60.94**	**56.22**	**54.95**
HELP	80.75	59.57	56.41	56.13	52.34	56.18	50.63	49.76	83.69	62.63	59.48	59.11	55.82	60.02	55.40	53.98
HELP_{HF}	**81.27**	**60.04**	**57.82**	**57.50**	**52.91**	**57.05**	**51.24**	**50.31**	**85.21**	**64.29**	**62.54**	**62.21**	**57.73**	**61.25**	**56.71**	**55.21**

Table 2. Classification accuracy under transfer-based black-box attacks. We use adversarial examples generated by *source model* to attack the *target model*.

Target Source	AT_{PGD}	AT_{HF}	TRADES	$\text{TRADES}_{\text{HF}}$
AT_{PGD}	52.64	61.50	62.76	63.61
AT_{HF}	60.12	56.39	62.35	63.32
TRADES	63.78	64.76	55.71	63.68
$\text{TRADES}_{\text{HF}}$	62.21	62.89	63.03	57.41



(a) Standard (b) AT_{PGD} (c) AT_{HF}

(d) early-stage AT_{PGD} (e) late-stage AT_{PGD} (f) AT_{HF}

Figure 5. The loss surfaces in the vicinity of an input are depicted in (a)-(c) for several models (standard model, AT_{PGD}, and AT_{HF}). These entail examining the direction of the gradient (d_{g}) and a direction chosen randomly (d_{r}) [8]. Additionally, in (d)-(f), the loss surfaces focus on the hider direction (d_{h}) and a random direction (d_{r}), which are depicted for different models (early-stage AT_{PGD} model, late-stage AT_{PGD} model, and AT_{HF}).

the existence of hiders. However, HFAT effectively suppresses the emergence of hiders as shown in Fig. 5 (f).

4.4. Ablation studies

4.4.1 Ablation study of auxiliary model

Although our method can identify potential areas for hiders by sampling the relative location information, we do not train the model directly using the samples computed by location ratio information as data augmentation, but use the auxiliary model to obtain the optimization direction. Therefore, in this subsection we compare the effects of utilizing auxiliary model and data augmentation. Additionally, we employ the Mixup [33] data augmentation method with a selected α value of 1.4 for further comparison. Tab. 3 reveals that although the direct inclusion of hiders leads to a substantial improvement in natural accuracy, it also results in a significant decrease in robust accuracy. Furthermore, the application of Mixup only yields marginal improvement. Nevertheless, via the implementation of the auxiliary model, HFAT significantly improves both its robustness and accuracy. We believe this is because samples computed by relative location information do not have significant attack performance under the current training epoch, and thus cannot be used directly for training to get good results.

Table 3. Comparison of robust and natural accuracies under different data augmentation methods. $\text{AT}_{\text{hiders}}$ includes hider samples directly into the training process as additional training data.

	Robust Accuracy	Natural Accuracy
AT_{PGD}	52.64	81.66
$\text{AT}_{\text{hiders}}$	48.90	85.73
Mixup	52.92	81.74
AT_{HF}	56.39	81.88

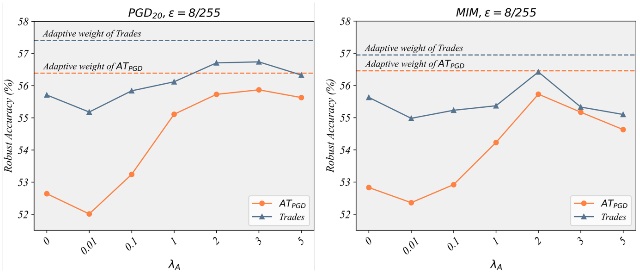


Figure 6. The robust accuracy of under white-box attack PGD_{20} and MIM are evaluated using various static auxiliary model weight settings: $\lambda_A = 0, 0.01, 0.1, 1, 2, 3,$ and 5. The dashed horizontal lines represent the accuracies of AT_{HF} and $\text{Trades}_{\text{HF}}$ respectively (i.e., using adaptive weighting mechanism). AT_{PGD} and Trades models are represented by distinct curves.

Table 4. Robust accuracy (evaluated by PGD_{20}), natural accuracy, and average runtime per epoch under different step value settings. We use AT_{PGD} as a reference.

	Robust Accuracy	Natural Accuracy	Training Time(s)
AT_{PGD}	52.64	81.66	192
step 1	53.52	83.89	211
step 3	54.26	82.75	248
step 5	56.39	81.88	287
step 7	56.47	81.13	326

4.4.2 Ablation study of adaptive weighting mechanism

We further investigate the impact of adaptive weighting λ on model performance. To convey concepts more effectively, we utilize λ_A to signify the weight of the auxiliary model branch, and λ_S to denote the weight of the standard adversarial training branch. In Fig. 6, we set the weight λ_S to a fixed value of 1, while adjusting λ_A of the auxiliary model to different values: $\lambda_A = 0.0, 0.01, 0.1, 1, 2, 3, 5$, and normalize the weights before the experiment. We employ PGD_{20} and MIM attacks to evaluate and represent the robust accuracy of adaptive weights using dashed lines. It is observed that the model's robustness gradually improves as the weight λ_A increases, reaching its peak when λ_A equals 2 or 3, followed by a decline. Additionally, the adaptive weight scheme exhibits significantly improved performance compared to static weights.

4.4.3 Ablation study of auxiliary model step

Introducing an auxiliary model incurs additional computational overhead, and to discuss this, we vary the step values for generating adversarial examples of the auxiliary model in Tab. 4. Larger step values require more time for computation. As the step value increases, both robust accuracy and natural accuracy gradually improve, reaching their optimal balance when the step value is set to 5. A noticeable decrease in natural accuracy is observed when the step size is set to 7. The reported training time includes the sequential computation for the auxiliary model and the standard adversarial training model.

4.5. Analysis of adaptive weighting mechanism

To provide evidence for the effectiveness of introducing an auxiliary model, we visualize the variations of adaptive weight values throughout the training process. Fig. 7 illustrates the evolution of the mean weight, λ_S, of the standard adversarial training model from around 0.5 in the early epochs to approximately 0.1 in the later stages of training. Conversely, the gradient weight λ_A of the auxiliary model gradually increases from 0.5 to 0.9. This observation indicates that as the training progresses, the gradient contribution of the auxiliary model becomes increasingly significant, highlighting the heightened importance of hiders' defense in the later stages of model training.

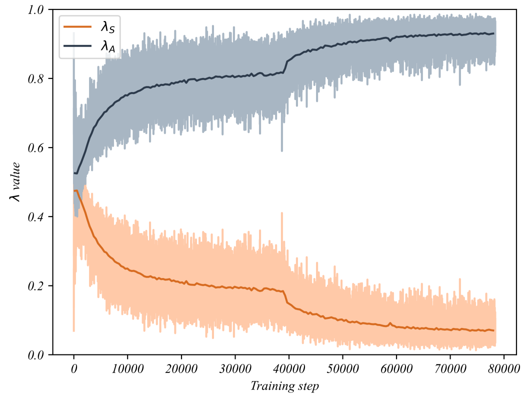


Figure 7. Visualization of adaptive weights as they vary with training steps. The blue hue signifies the weights λ_A of the auxiliary model, whereas the orange hue corresponds to the weights λ_S of the standard adversarial training branch. In addition, the darker line represents the average value for each epoch.

5. Future work

We identify hiders as hidden high-risk areas and propose HFAT that defends against both hiders and adversarial examples using iterative evolution optimization strategy and auxiliary model. Hider detection in this study uses empirical fitting with Gaussian distribution. We encourage introducing an evaluation metric to assess the model's capability in detecting hidden threats, which could advance research on self-supervision-based methods for hider prevention.

6. Conclusion

This paper highlights the limitations of conventional adversarial training, which focuses solely on worst-case adversarial examples and neglects the hidden threats in secure regions. We introduce "hiders" samples that can be defended or correctly classified initially but are vulnerable later. Our method, HFAT, uses an auxiliary model to reveal potential threats and offer optimized guidance for enhancing robustness in the regions prone to hider emergence. HFAT improves model robustness and accuracy through joint optimization and adaptive weighting. Our experiments not only show that HFAT can provide stronger robustness and better accuracy, but also demonstrate HFAT's effectiveness in mitigating hider-related risks.

7. Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62106116, 62276149, U2341228), China Meteorological Administration under Grant QBZ202316, Natural Science Foundation of Ningbo of China (No. 2023J027), as well as by the High Performance Computing Centers at Eastern Institute of Technology, Ningbo, and Ningbo Institute of Digital Twin.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pages 484–501. Springer, 2020. 6

[2] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 4

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 3, 6

[4] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 3, 6

[5] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pages 2196–2205. PMLR, 2020. 6

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 1, 3, 6

[8] Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. Adversarial distributional training for robust deep learning. *Advances in Neural Information Processing Systems*, 33:8270–8283, 2020. 2, 6, 7

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 6

[10] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020. 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 6

[13] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Las-at: adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13398–13408, 2022. 4

[14] Gaojie Jin, Xinping Yi, Dengyu Wu, Ronghui Mu, and Xiaowei Huang. Randomized adversarial training via taylor expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16447–16457, 2023. 1

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[16] Qian Li, Yuxiao Hu, Ye Liu, Dongxiao Zhang, Xin Jin, and Yuntian Chen. Discrete point-wise attack is not enough: Generalized manifold adversarial attack for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20575–20584, 2023. 1

[17] Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, and Sabine Süsstrunk. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. *Advances in Neural Information Processing Systems*, 33:21476–21487, 2020. 6

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2, 3, 4, 6

[19] Marie-Hélène Masson and Thierry Denoeux. Inferring a possibility distribution from empirical data. *Fuzzy sets and systems*, 157(3):319–340, 2006. 4

[20] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6

[21] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*, 2021. 2, 4, 6

[22] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019. 4

[23] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020. 3

[24] Naoki Saito, Ronald R Coifman, Frank B Geshwind, and Fred Warner. Discriminant feature extraction using empirical probability density estimation and a local basis library. *Pattern Recognition*, 35(12):2841–2852, 2002. 4

[25] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020. 6

[26] Aris Spanos. *Probability theory and statistical inference: Empirical modeling with observational data*. Cambridge University Press, 2019. 4

[27] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 6

[28] Yisen Wang, Xuejiao Deng, Songbai Pu, and Zhiheng Huang. Residual convolutional ctc networks for automatic speech recognition. *arXiv preprint arXiv:1702.07793*, 2017. 1

[29] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019. 2, 4, 6

[30] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020. 2, 3, 4, 6

[31] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019. 6

[32] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6

[33] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 7

[34] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 2, 3, 4, 6