Scaling Adversarial Training to Large Perturbation Bounds

Sravanti Addepalli*, Samyak Jain*, Gaurang Sriramanan, and R.Venkatesh Babu

Video Analytics Lab, Department of Computational and Data Sciences, Indian Institute of Science, Bangalore

Abstract. The vulnerability of Deep Neural Networks to Adversarial Attacks has fuelled research towards building robust models. While most Adversarial Training algorithms aim at defending attacks constrained within low magnitude Lp norm bounds, real-world adversaries are not limited by such constraints. In this work, we aim to achieve adversarial robustness within larger bounds, against perturbations that may be perceptible, but do not change human (or Oracle) prediction. The presence of images that flip Oracle predictions and those that do not makes this a challenging setting for adversarial robustness. We discuss the ideal goals of an adversarial defense algorithm beyond perceptual limits, and further highlight the shortcomings of naively extending existing training algorithms to higher perturbation bounds. In order to overcome these shortcomings, we propose a novel defense, Oracle-Aligned Adversarial Training (OA-AT), to align the predictions of the network with that of an Oracle during adversarial training. The proposed approach achieves state-of-the-art performance at large epsilon bounds (such as an L-inf bound of 16/255 on CIFAR-10) while outperforming existing defenses (AWP, TRADES, PGD-AT) at standard bounds (8/255) as well.

1 Introduction

Deep Neural Networks are known to be vulnerable to Adversarial Attacks, which are perturbations crafted with an intention to fool the network [27]. With the rapid increase in deployment of Deep Learning algorithms in various critical applications such as autonomous navigation, it is becoming increasingly crucial to improve the Adversarial robustness of these models. In a classification setting, Adversarial attacks can flip the prediction of a network to even unrelated classes, while causing no change in a human's prediction (Oracle label).

The definition of adversarial attacks involves the prediction of an Oracle, making it challenging to formalize threat models for the training and verification of adversarial defenses. The widely used convention that overcomes this challenge is the ℓ_p norm based threat model with low-magnitude bounds to ensure imperceptibility [10,3]. For example, attacks constrained within an ℓ_∞ norm

* Equal contribution.
Correspondence to: Sravanti Addepalli <sravantia@iisc.ac.in>

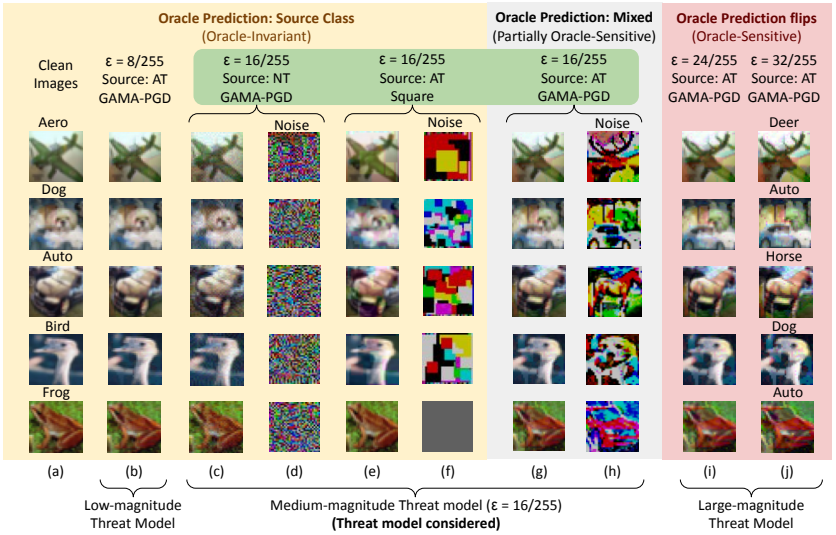

Fig. 1. **Perturbations within different threat models:** Adversarial images (b, c, e, g, i, j) and perturbations (d, f, h) along with the corresponding clean image (a) for various ℓ_∞ norm bounds on CIFAR-10. Attacks are generated from an Adversarially Trained model (AT) or a Normally Trained model (NT) using the gradient-based attack GAMA-PGD [25] or the Random-search based attack Square [1]. The medium-magnitude threat model is challenging since it consists of attacks which are Oracle-Invariant and partially Oracle-Sensitive.

of 8/255 on the CIFAR-10 dataset are imperceptible to the human eye as shown in Fig.1(b), ensuring that the Oracle label is unchanged. The goal of Adversarial Training within such a threat model is to ensure that the prediction of the model is consistent within the considered perturbation radius ε, and matches the label associated with the unperturbed image.

While low-magnitude ℓ_p norm based threat models form a crucial subset of the widely accepted definition of adversarial attacks [9], they are not sufficient, as there exist valid attacks at higher perturbation bounds as well, as shown in Fig.1(c) and (e). However, the challenge at large perturbation bounds is the existence of attacks that can flip Oracle labels as well [28], as shown in Fig.1(g), (i) and (j). Naively scaling existing Adversarial Training algorithms to large perturbation bounds would enforce consistent labels on images that flip the Oracle prediction as well, leading to a conflict in the training objective as shown in Fig.2. This results in a large drop in clean accuracy, as shown in Table-1. This has triggered interest towards developing perceptually aligned threat models, and defenses that are robust under these settings [17]. However, finding a perceptually aligned metric is as challenging as building a network that can replicate oracle predictions [28]. Thus, it is crucial to investigate adversarial robustness using the well-defined ℓ_p norm metric under larger perturbation bounds.

Table 1. **CIFAR-10: Standard Adversarial Training using Large-ε:** Performance (%) of various existing Defenses trained using $\varepsilon = 8/255$ or $16/255$ against attacks bound within $\varepsilon = 8/255$ and $16/255$. A large drop in clean accuracy is observed with existing approaches [33,30,18,34] when trained using perturbations with $\varepsilon = 16/255$.

Method	Attack ε (Training)	Clean Acc	GAMA (8/255)	AA (8/255)	GAMA (16/255)	Square (16/255)
TRADES	8/255	80.53	49.63	49.42	19.27	27.82
TRADES	16/255	75.30	35.64	35.12	10.10	18.87
AWP	8/255	80.47	**50.06**	**49.87**	19.66	28.51
AWP	16/255	71.63	40.85	40.55	15.92	24.16
PGD-AT	8/255	81.12	49.03	48.58	15.77	26.47
PGD-AT	16/255	64.93	46.66	46.21	**26.73**	**32.25**
FAT	8/255	**84.36**	48.41	48.14	15.18	25.07
FAT	16/255	75.27	47.68	47.34	22.93	29.47

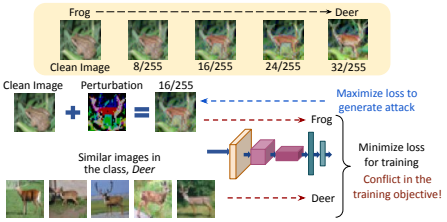

Fig. 2. **Issues with Standard Adversarial Training at Large-ε:** An adversarial example generated from the original image of a frog looks partially like a deer at an ℓ_∞ bound of $16/255$, but is trained to predict the true label, Frog. This induces a conflicting objective, leading to a large drop in clean accuracy.

In this work, we aim to improve robustness at larger epsilon bounds, such as an ℓ_∞ norm bound of $16/255$ on the CIFAR-10 and CIFAR-100 datasets [16].

We define this as a moderate-magnitude bound, and discuss the ideal goals for achieving robustness under this threat model in Sec.3.3. We further propose a novel defense Oracle-Aligned Adversarial Training (OA-AT), which attempts to align the predictions of the network with that of an Oracle, rather than enforcing all samples within the constraint set to have the same label as the original image. Our contributions have been summarized below:

- We propose Oracle-Aligned Adversarial Training (OA-AT) to improve robustness within the defined moderate-ε threat model.
- We demonstrate superior performance when compared to state-of-the-art methods such as AWP [30], TRADES [33] and PGD-AT [18] at $\varepsilon = 16/255$ while also performing better at $\varepsilon = 8/255$ on CIFAR-10 and SVHN. We also demonstrate improved performance on challenging datasets such as CIFAR-100 and Imagenette (10-class subset of Imagenet with 160×160 images).
- We achieve improvements over the baselines even at larger model capacities such as WideResNet-34-10, and demonstrate results that outperform existing methods on the RobustBench leaderboard.
- We show the relation between contrast level of images and the existence of attacks that can flip the Oracle label within a given perturbation bound, and use this observation for constructing better evaluation metrics at large perturbation bounds.

Our code is available here: https://github.com/val-iisc/OAAT.

2 Related Works

Robustness against imperceptible attacks: Following the discovery of adversarial examples by Szegedy et al., [27], a myriad of adversarial attack and defense methods have been proposed. Adversarial Training has emerged as the most successful defense strategy against ℓ_p norm bound imperceptible attacks. PGD Adversarial Training (PGD-AT) proposed by Madry et al. [18] constructs multi-step adversarial attacks by maximizing Cross-Entropy loss within the considered threat model and subsequently minimizes the same for training.

This was followed by several adversarial training methods [33,34,22,30,25,20] that improved accuracy against such imperceptible threat models further.

Zhang et al. [33] proposed the TRADES defense, which maximizes the Kullback-Leibler (KL) divergence between the softmax outputs of adversarial and clean samples for attack generation, and minimizes the same in addition to the Cross-Entropy loss on clean samples for training.

Improving Robustness of base defenses: Wu et al. [30] proposed an additional step of Adversarial Weight Perturbation (AWP) to maximize the training loss, and further train the perturbed model to minimize the same. This generates a flatter loss surface [26], thereby improving robust generalization. While this can be integrated with any defense, AWP-TRADES is the state-of-the-art adversarial defense today.

On similar lines, the use of stochastic weight averaging of model weights [15] is also seen to improve the flatness of loss surface, resulting in a boost in adversarial robustness [11,5]. Recent works attempt to use training techniques such as early stopping [22], optimal weight decay [20], Cutmix data augmentation [31,21] and label smoothing [21] to achieve enhanced robust performance on base defenses such as PGD-AT [18] and TRADES [33]. We utilize some of these methods in our approach (Sec.7), and also present improved baselines by combining AWP-TRADES [30] with these enhancements.

Robustness against large perturbation attacks: Shaeiri et al. [23] demonstrate that the standard formulation of adversarial training is not well-suited for achieving robustness at large perturbations, as the loss saturates very early. The authors propose Extended Adversarial Training (ExAT), where a model trained on low-magnitude perturbations ($\varepsilon = 8/255$) is fine-tuned with large magnitude perturbations ($\varepsilon = 16/255$) for just 5 training epochs, to achieve improved robustness at large perturbations. The authors also discuss the use of a varying epsilon schedule to improve training convergence. Friendly Adversarial Training (FAT) [34] performs early-stopping of an adversarial attack by thresholding the number of times the model misclassifies the image during attack generation. The threshold is increased over training epochs to increase the strength of the attack over training. Along similar lines, Sitawarin et al. [24] propose Adversarial Training with Early Stopping (ATES), which performs early stopping of a PGD attack based on the margin (difference between true and maximum probability class softmax outputs) of the perturbed image being greater than a threshold that is increased over epochs. We compare against these methods and improve upon them significantly using our proposed approach (Sec.4).

3 Preliminaries and Threat Model

3.1 Notation

We consider an N-class image classification problem with access to a labelled training dataset $\mathcal{D}$. The input images are denoted by $x \in \mathcal{X}$ and their corresponding labels are denoted as $y \in \{1, ..., N\}$. The function represented by the Deep Neural Network is denoted by f_θ where $\theta \in \Theta$ denotes the set of network parameters. The N-dimensional softmax output of the input image x is denoted as $f_\theta(x)$. Adversarial examples are defined as images that are crafted specifically to fool a model into making an incorrect prediction [9]. An adversarial image corresponding to a clean image x would be denoted as $\widetilde{x}$. The set of all images within an ℓ_p norm ball of radius ε is defined as $\mathcal{S}(x) = \{\hat{x} : ||\hat{x} - x||_p < \varepsilon\}$.

In this work, we specifically consider robustness to ℓ_∞ norm bound adversarial examples. We define the Oracle prediction of a sample x as the label that a human is likely to assign to the image, and denote it as $O(x)$. For a clean image, $O(x)$ would correspond to the true label y, while for a perturbed image it could differ from the original label.

3.2 Nomenclature of Adversarial Attacks

Tramer et al. [28] discuss the existence of two types of adversarial examples: Sensitivity-based examples, where the model prediction changes while the Oracle prediction remains the same as the unperturbed image, and Invariance-based examples, where the Oracle prediction changes while the model prediction remains unchanged. Models trained using standard empirical risk minimization are susceptible to sensitivity-based attacks, while models which are overly robust to large perturbation bounds could be susceptible to invariance-based attacks. Since these definitions are model-specific, we define a different nomenclature which only depends on the input image and the threat model considered:

- Oracle-Invariant set $OI(x)$ is defined as the set of all images within the bound $\mathcal{S}(x)$, that preserve Oracle label. Oracle is invariant to such attacks:

$$OI(x) := \{\hat{x} : O(\hat{x}) = O(x), \hat{x} \in \mathcal{S}(x)\} \tag{1}$$

- Oracle-Sensitive set $OS(x)$ is defined as the set of all images within the bound $\mathcal{S}(x)$, that flip the Oracle label. Oracle is sensitive to such attacks:

$$OS(x) := \{\hat{x} : O(\hat{x}) \neq O(x), \hat{x} \in \mathcal{S}(x)\} \tag{2}$$

3.3 Objectives of the Proposed Defense

Defenses based on the conventional ℓ_p norm threat model attempt to train models which are invariant to all samples within $\mathcal{S}(x)$. This is an ideal requirement for low ε-bound perturbations, where the added noise is imperceptible, and hence all samples within the threat model are Oracle-Invariant. An example of a low

ε-bound constraint set is the ℓ_∞ threat model with $\varepsilon = 8/255$ for the CIFAR-10 dataset, which produces adversarial examples that are perceptually similar to the corresponding clean images, as shown in Fig.1(b).

As we move to larger ε bounds, Oracle-labels begin to change, as shown in Fig.1(g, i, j). For a very high perturbation bound such as 32/255, the changes produced by an attack are clearly perceptible and in many cases flip the Oracle label as well. Hence, robustness at such large bounds is not of practical relevance. The focus of this work is to achieve robustness within a moderate-magnitude ℓ_p norm bound, where some perturbations look partially modified (Fig.1(g)), while others look unchanged (Fig.1(c, e)), as is the case with $\varepsilon = 16/255$ for CIFAR-10. The existence of attacks that do not significantly change the perception of the image necessitates the requirement of robustness within such bounds, while the existence of partially Oracle-Sensitive samples makes it difficult to use standard adversarial training methods on the same. The ideal goals for training defenses under this moderate-magnitude threat model are described below:

- Robustness against samples which belong to $OI(x)$
- Sensitivity towards samples which belong to $OS(x)$, with model's prediction matching the Oracle label
- No specification on samples which cannot be assigned an Oracle label.

Given the practical difficulty in assigning Oracle labels during training and evaluation, we consider the following subset of these ideal goals in this work:

- Robustness-Accuracy trade-off, measured using accuracy on clean samples and robustness against valid attacks within the threat model
- Robustness against all attacks within an imperceptible radius ($\varepsilon = 8/255$ for CIFAR-10), measured using strong white-box attacks [7,25]
- Robustness to Oracle-Invariant samples within a larger radius ($\varepsilon = 16/255$ for CIFAR-10), measured using gradient-free attacks [1,4]

4 Proposed Method

In order to achieve the goals discussed in Sec.3.3, we require to generate Oracle-Sensitive and Oracle-Invariant samples and impose specific training losses on each of them individually. Since labeling adversarial samples as Oracle-Invariant or Oracle-Sensitive is expensive and cannot be done while training networks, we propose to use attacks which ensure a given type of perturbation (OI or OS) by construction, and hence do not require explicit annotation.

Generation of Oracle-Sensitive examples: Robust models are known to have perceptually aligned gradients [29]. Adversarial examples generated using a robust model tend to look like the target (other) class images at large perturbation bounds, as seen in Fig.1(g, i, j). We therefore use large ε-bound white-box adversarial examples generated from the model being trained as Oracle-Sensitive samples, and the model prediction as a proxy to the Oracle prediction.

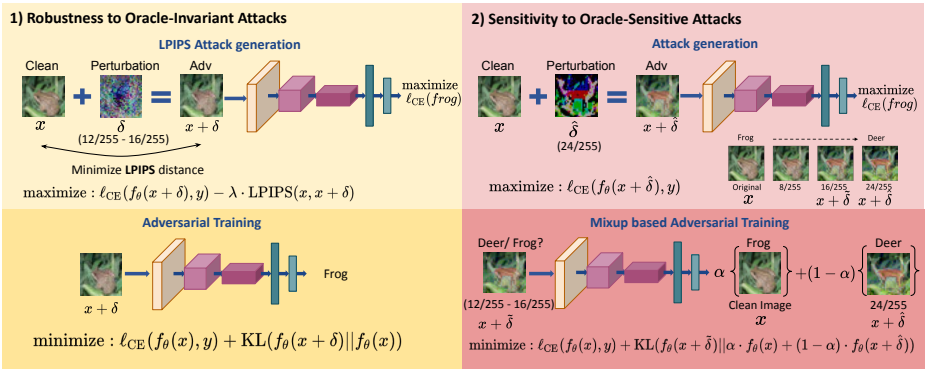

Fig. 3. **Oracle-Aligned Adversarial Training:** The proposed defense OA-AT involves alternate training on Oracle-Invariant and Oracle-Sensitive samples. 1) Oracle-Invariant samples are generated by minimizing the LPIPS distance between the clean and perturbed images in addition to the maximization of the Classification Loss. 2) Oracle-Sensitive samples are trained using a convex combination of the predictions of the clean image and the perturbed image at a larger perturbation bound as reference in the KL divergence loss.

Generation of Oracle-Invariant examples: While the strongest Oracle-Invariant examples are generated using the gradient-free attacks Square [1] and Ray-S [4], they require a large number of queries (5000 to 10000), which is computationally expensive for use in adversarial training. Furthermore, reducing the number of queries weakens the attack significantly. The most efficient attack that is widely used for adversarial training is the PGD 10-step attack. However, it cannot be used for the generation of Oracle-Invariant samples as gradient-based attacks generated from adversarially trained models produce Oracle-Sensitive samples. We propose to use the Learned Perceptual Image Patch Similarity (LPIPS) measure for the generation of Oracle-Invariant attacks, as it is known to match well with perceptual similarity based on a study involving human annotators [35,17]. Further, we observe that while the standard AlexNet model used in prior work [17] fails to distinguish between Oracle-Invariant and Oracle-Sensitive samples, an adversarially trained model is able to distinguish between the two effectively (Ref: Fig.3 in the Supplementary). We therefore propose to minimize the LPIPS distance between natural and perturbed images, in addition to the maximization of Cross-Entropy loss for attack generation: $\mathcal{L}_{CE}(x, y) - \lambda \cdot \text{LPIPS}(x, \hat{x})$. The ideal setting of λ is the minimum value that transforms attacks from Oracle-Sensitive to Oracle-Invariant (OI) for majority of the images. This results in the generation of strong Oracle-Invariant (OI) attacks. We present several Oracle-Invariant examples for visual inspection in Fig.5 in Supplementary.

Oracle-Aligned Adversarial Training (OA-AT): The training algorithm for the proposed defense, Oracle-Aligned Adversarial Training (OA-AT) is presented in Algorithm-1 and illustrated in Fig.3. We denote the maximum pertur-

Algorithm 1 Oracle-Aligned Adversarial Training

1: **Input:** Deep Neural Network f_θ with parameters θ, Training Data $\{x_i, y_i\}_{i=1}^{M}$, Epochs T, Learning Rate η, Perturbation budget ε_{max}, Adversarial Perturbation function $A(x, y, \ell, \varepsilon)$ which maximises loss ℓ
2: **for** epoch $= 1$ **to** T **do**
3: $\quad\widetilde{\varepsilon} = \max\{\varepsilon_{max}/4, \varepsilon_{max} \cdot \text{epoch}/T\}$
4: $\quad$**for** $i = 1$ **to** M **do**
5: $\quad\quad\delta_i \sim U(-\min(\widetilde{\varepsilon}, \varepsilon_{max}/4), \min(\widetilde{\varepsilon}, \varepsilon_{max}/4))$
6: $\quad\quad$**if** $\widetilde{\varepsilon} < 3/4 \cdot \varepsilon_{max}$ **then**
7: $\quad\quad\quad\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i)$, $\widetilde{\delta}_i = A(x_i, y_i, \ell, \widetilde{\varepsilon})$
8: $\quad\quad\quad L_{adv} = \text{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) || f_\theta(x_i)\right)$
9: $\quad\quad$**else if** $i \,\%\, 2 = 0$ **then**
10: $\quad\quad\quad\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i)$, $\widehat{\delta}_i = A(x_i, y_i, \ell, \varepsilon_{ref})$, $\widetilde{\delta}_i = \Pi_\infty(\widehat{\delta}_i, \widetilde{\varepsilon})$
11: $\quad\quad\quad L_{adv} = \text{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) \,||\, \alpha \cdot f_\theta(x_i) + (1-\alpha) \cdot f_\theta(x_i + \widehat{\delta}_i)\right)$
12: $\quad\quad$**else**
13: $\quad\quad\quad\delta_i \sim U(-\widetilde{\varepsilon}, \widetilde{\varepsilon})$
14: $\quad\quad\quad\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i) - \text{LPIPS}(x_i, x_i + \delta_i)$, $\widetilde{\delta}_i = A(x_i, y_i, \ell, \widetilde{\varepsilon})$
15: $\quad\quad\quad L_{adv} = \text{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) \,||\, f_\theta(x_i)\right)$
16: $\quad\quad L = \ell_{CE}(f_\theta(x_i), y_i) + L_{adv}$
17: $\quad\quad\theta = \theta - \eta \cdot \nabla_\theta L$

bation bound used for attack generation during the training by ε_{max}. We use the AWP-TRADES formulation [33,30] as the base implementation. Similar to Wu et al. [30], we use 10 steps of optimization for attack generation and one additional weight perturbation step. We maximize the classification loss on $x_i + 2 \cdot \widetilde{\delta}_i$ (where $\widetilde{\delta}_i$ is the attack) in the additional weight perturbation step (instead of $x_i + \widetilde{\delta}_i$ [30]), in order to achieve better smoothness in the loss surface. We start training with attacks constrained within a perturbation bound of $\varepsilon_{max}/4$ upto one-fourth the training epochs (Alg.1, L6-L8), and ramp up this value linearly to ε_{max} at the last epoch alongside a cosine learning rate schedule. The use of a fixed epsilon initially helps in improving the adversarial robustness faster, while the use of an increasing epsilon schedule later results in better training stability [23]. We use 5 attack steps upto $\varepsilon_{max}/4$ to reduce computation and 10 attack steps later.

We perform standard adversarial training upto a perturbation bound of $3/4 \cdot \varepsilon_{max}$ as the attacks in this range are imperceptible, based on the chosen moderate-magnitude threat model discussed in Sec.3.3. Beyond this, we start incorporating separate training losses for Oracle-Invariant and Oracle-Sensitive samples in alternate training iterations (Alg.1, L9-L15), as shown in Fig.3. Oracle-Sensitive samples are generated by maximizing the classification loss in a PGD attack formulation. Rather than enforcing the predictions of such attacks to be similar to the original image, we allow the network to be partially sensitive to such attacks by training them to be similar to a convex combination of predictions on the clean image and perturbed samples constrained within a bound of ε_{ref}, which is chosen to be greater than or equal to ε_{max} (Alg.1, L10).

This component of the overall training loss is shown below:

$$KL\big(f_\theta(x_i+\widetilde{\delta}_i)\ ||\ \alpha\ f_\theta(x_i)+(1-\alpha)\ f_\theta(x_i+\widehat{\delta}_i)\big) \tag{3}$$

Here $\widetilde{\delta}_i$ is the perturbation at the varying epsilon value $\widetilde{\varepsilon}$, and $\widehat{\delta}_i$ is the perturbation at ε_{ref}. This loss formulation results in better robustness-accuracy trade-off as shown in E1 versus E3 of Table-4. In the alternate iteration, we use the LPIPS metric to efficiently generate strong Oracle-Invariant attacks during training (Alg.1, L14). We perform exponential weight-averaging of the network being trained and use this for computing the LPIPS metric for improved and stable results (E1 versus E2 and F1 versus F2 in Table-4). We therefore do not need additional training or computation time for training this model. We increase α and λ over training, as the nature of attacks changes with varying $\widetilde{\varepsilon}$. The use of both Oracle-Invariant (OI) and Oracle-Sensitive (OS) samples ensures robustness to Oracle-Invariant samples while allowing sensitivity to partially Oracle-Sensitive samples.

5 Analysing Oracle Alignment of Adversarial Attacks

We first consider the problem of generating Oracle-Invariant and Oracle-Sensitive attacks in a simplified, yet natural setting to enable more fine-grained theoretical analysis. We consider a binary classification task as introduced by Tsipras et al. [29], consisting of data samples (x, y), with $y \in \{+1, -1\}$, $x \in \mathbb{R}^{d+1}$. Further,

$$x_1 = \begin{cases} y, & \text{w.p. } p \\ -y, & \text{w.p. } 1-p \end{cases}, \ x_i \sim \mathcal{N}(\alpha y, 1)\ \forall i \in \{2, \ldots, d+1\}$$

In this setting, x_1 can be viewed as a feature that is strongly correlated with the Oracle Label y when the Bernoulli parameter p is sufficiently large (for eg: $p \approx 0.90$), and thus corresponds to an Oracle Sensitive feature. On the other hand, $x_2, \ldots, x_{d+1}$ are spurious features that are positively correlated (in a weak manner) to the Oracle label y, and are thus Oracle Invariant features. Building upon theoretical analysis presented by Tsipras et al. [29], we make a series of observations, whose details we expound in the Supplementary Section 2:

Observation 1. Adversarial perturbations of a standard, non-robust classifier utilize spurious features, resulting in Oracle Invariant Samples that are weakly anti-correlated with the Oracle label y.

Observation 2. Adversarial perturbations of a robust model result in Oracle Sensitive Samples, utilizing features strongly correlated with the Oracle label y.

6 Role of Image Contrast in Robust Evaluation

As shown in Fig.1, perturbations constrained within a low-magnitude bound (Fig.1(b)) do not change the perceptual appearance of an image, whereas perturbations constrained within very large bounds such as $\varepsilon = 32/255$ (Fig.1(j))

(a) **CIFAR-10** (b) **SVHN**

Fig. 4. **Relation between the contrast level of an image and the Oracle-Sensitivity of adversarial examples** within a given perturbation bound. First and second rows show low contrast images, and third and fourth rows show high contrast images. Column (a) shows the original clean image and columns (b-e) show adversarial examples at different perturbation bounds generated at the largest bound in (e) and projected to the other bounds in (b, c, d). The adversarial perturbation is shown in column (f). Adversarial examples in columns (d) and (e) are Oracle-Invariant for the high contrast images, and Oracle-Sensitive for the low contrast images.

flip the Oracle prediction. As noted by Balaji et al. [2], the perturbation radius at which the Oracle prediction changes varies across images. We hypothesize that the contrast level of an image plays an important role in determining the minimum perturbation magnitude ε_{OS} that can flip the Oracle prediction of an image to generate an Oracle-Sensitive (OS) sample. We visualize a few High-Contrast and Low-Contrast images of the CIFAR-10 and SVHN datasets in Fig.4 (more comprehensive visualisations are made available in Fig.10-15 in the Supplementary). We observe that High-contrast (HC) images are Oracle-Invariant even at large perturbation bounds, while Low-Contrast (LC) images are Oracle-Sensitive at lower perturbation bounds as well. Based on this, we present robust evaluations at large epsilon bounds on images of varying contrast levels in Fig.5.

7 Experiments and Results

In this section, we present detailed robust evaluations of the proposed approach along with various existing defenses on the CIFAR-10 [16], CIFAR-100 [16], SVHN [19] and Imagenette [14] datasets. We report adversarial robustness against the strongest known attacks, AutoAttack (AA) [7] and GAMA PGD-100 (GAMA) [25] for $\varepsilon = 8/255$ in order to obtain the worst-case robust accuracy. For larger bounds such as 12/255 and 16/255, we primarily aim for robustness against an ensemble of the Square [1] and Ray-S [4] attacks, as they generate strong Oracle-Invariant examples. On the SVHN dataset, we find that the perturbation bound for imperceptible attacks is $\varepsilon = 4/255$, and consider robustness within 12/255 (Fig.10, 11 in the Supplementary).

Table 2. **Comparison with existing methods:** Performance (%) of the proposed defense OA-AT when compared to baselines against the attacks, GAMA-PGD100 [25], AutoAttack (AA) [7] and an ensemble of Square [1] and Ray-S [4] attacks (SQ+RS), with different ε bounds. Sorted by AutoAttack (AA) accuracy at $\varepsilon = 8/255$ for CIFAR-10, CIFAR-100 and Imagenette, and $4/255$ for SVHN.

(a) **CIFAR-10, SVHN**

Method	Metrics of interest				Others	
	Clean	GAMA 8/255	AA 8/255	SQ+RS 16/255	GAMA 16/255	AA 16/255
CIFAR-10 (ResNet-18), 110 epochs						
FAT	**84.36**	48.41	48.14	23.22	15.18	14.22
PGD-AT	79.38	49.28	48.68	25.43	18.18	17.00
AWP	80.32	49.06	48.89	25.99	19.17	18.77
ATES	80.95	49.57	49.12	26.43	18.36	16.30
TRADES	80.53	49.63	49.42	26.20	19.27	18.23
ExAT + PGD	80.68	50.06	49.52	25.13	17.81	19.53
ExAT + AWP	80.18	49.87	49.69	27.04	20.04	16.67
AWP	80.47	50.06	49.87	27.20	19.66	19.23
Ours	80.24	**51.40**	**50.88**	**29.56**	**22.73**	**22.05**
CIFAR-10 (ResNet-34), 110 epochs						
AWP	83.89	52.64	52.44	27.69	20.23	19.69
OA-AT (Ours)	**84.07**	**53.54**	**53.22**	**30.76**	**22.67**	**22.00**
CIFAR-10 (WRN-34-10), 200 epochs						
AWP	85.36	56.34	56.17	30.87	23.74	23.11
OA-AT (Ours)	**85.32**	**58.48**	**58.04**	**35.31**	**26.93**	**26.57**
SVHN (PreActResNet-18), 110 epochs						
Method	**Clean**	**GAMA 4/255**	**AA 4/255**	**SQ+RS 12/255**	**GAMA 12/255**	**AA 12/255**
AWP	91.91	75.92	75.72	35.49	30.70	30.31
OA-AT (Ours)	**94.61**	**78.37**	**77.96**	**39.24**	**34.25**	**33.63**

(b) **CIFAR-100, ImageNette**

Method	Metrics of interest				Others	
	Clean	GAMA 8/255	AA 8/255	SQ+RS 16/255	GAMA 16/255	AA 16/255
CIFAR-100 (ResNet-18), 110 epochs						
AWP	58.81	25.51	25.30	11.39	8.68	8.29
AWP+	59.88	25.81	25.52	11.85	8.72	8.28
OA-AT (no LS)	60.27	26.41	26.00	13.48	**10.47**	**9.95**
OA-AT (Ours)	**61.70**	**27.09**	**26.77**	**13.87**	10.40	9.91
CIFAR-100 (PreActResNet-18), 200 epochs						
AWP	58.85	25.58	25.18	11.29	8.63	8.19
AWP+	**62.11**	26.21	25.74	12.23	9.21	8.55
OA-AT (Ours)	62.02	**27.45**	**27.14**	**14.52**	**10.64**	**10.10**
CIFAR-100 (WRN-34-10), 110 epochs						
AWP	62.41	29.70	29.54	14.25	11.06	10.63
AWP+	62.73	29.92	29.59	14.96	11.55	11.04
OA-AT (no LS)	65.22	30.75	**30.35**	16.77	12.65	11.95
OA-AT (Ours)	**65.73**	**30.90**	**30.35**	**17.15**	**13.21**	**12.01**
Imagenette (ResNet-18), 110 epochs						
Method	**Clean**	**GAMA 8/255**	**AA 8/255**	**SQ+RS 16/255**	**GAMA 16/255**	**AA 16/255**
AWP	82.73	57.52	57.40	42.52	29.14	28.86
OA-AT (Ours)	**82.98**	**59.51**	**59.31**	**48.01**	**48.66**	**31.78**

For each baseline on CIFAR-10, we find the best set of hyperparameters to achieve clean accuracy of around 80% to ensure a fair comparison across all methods. We further perform baseline training across various ε values and report the best in Table-2a. We note that existing defenses do not perform well when trained using large ε bounds such as 16/255 as shown in Table-1 (more detailed results available in Table-2,3 in Supplementary). On other datasets, we present comparative analysis primarily with AWP [30], the leading defense amongst prior methods on the RobustBench Leaderboard [6] in the setting without additional or synthetic training data, which we consider in this work. We further compare the proposed approach with the AWP baseline using various model architectures (ResNet-18, ResNet-34 [12], WideResNet-34-10 [32] and PreActResNet-18 [13]).

Contrary to prior works [22,21], we obtain additional gains with the use of the augmentation technique, AutoAugment [8]. We also use Model Weight Averaging (WA) [15,11,5] to obtain better generalization performance, especially at larger model capacities. To ensure a fair comparison, we use these methods to obtain improved baselines as well, and report this as AWP+ in Table-2 if any improvement is observed (more comprehensive results in Sec.7.4 of the Supplementary). As observed by Rebuffi et al. [21], we find that label-smoothing and the use of warmup in the learning rate scheduler helps achieve an additional boost in robustness. However, we report our results without including this as well (no LS) to highlight the gains of the proposed method individually.

Table 3. **Comparison with RobustBench Leaderboard [6] Results:** Performance (%) of the proposed method (OA-AT) when compared to AWP [30], which is the state-of-the-art amongst methods that do not use additional training data/ synthetic data on the RobustBench Leaderboard.

Method	Clean Acc	ℓ_∞ (AA) 8/255	ℓ_∞ (OI) 16/255	ℓ_2 (AA) $\varepsilon = 0.5$	ℓ_2 (AA) $\varepsilon = 1$	ℓ_1 (AA) $\varepsilon = 5$	ℓ_0 (PGD_0) $\varepsilon = 7$	Comm Corr
CIFAR-10 (WRN-34-10)								
AWP	**85.36**	56.17	30.87	60.68	28.86	37.29	39.09	75.83
Ours	85.32	**58.04**	**35.31**	**64.08**	**34.54**	**45.72**	**44.40**	**76.78**
CIFAR-100 (WRN-34-10)								
AWP	62.73	29.59	14.96	36.62	17.05	21.88	17.40	50.73
Ours	**65.73**	**30.35**	**17.15**	**37.21**	**17.41**	**25.75**	**29.20**	**54.88**

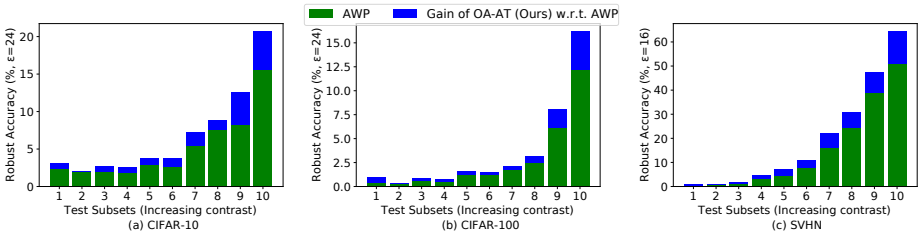

Fig. 5. **Evaluation across test subsets of increasing contrast levels:** Here we plot the gain in robust accuracy of the proposed defense OA-AT over AWP [30]. The proposed defense achieves higher gains as contrast increases, verifying that the proposed approach is more robust to the Oracle-Invariant white-box attacks on High-Contrast images.

From Table-2, we observe that the proposed defense achieves significant and consistent gains across all metrics specified in Sec.3.3. The proposed approach outperforms existing defenses by a significant margin on all four datasets, over different network architectures. Although we train the model for achieving robustness at larger ε bounds, we achieve an improvement in the robustness at the low ε bound (such as $\varepsilon = 8/255$ on CIFAR-10) as well, which is not observed in any existing method (Sec.7 of Supplementary). We also report the results on ℓ_2 Norm adversaries in Table-4 of Supplementary. As shown in Fig.5, the proposed defense achieves higher gains on the high contrast test subsets of different datasets, verifying that the proposed approach has better robustness against Oracle-Invariant attacks, and not against Oracle-Sensitive attacks.

RobustBench Leaderboard Comparisons: As shown in Table-3, using the proposed method, we obtain a significant improvement over state-of-the-art results reported on the RobustBench Leaderboard (AWP) without the use of additional/ synthetic data on both CIFAR-10 and CIFAR-100 datasets. We observe that the proposed approach achieves significant gains against ℓ_∞ norm

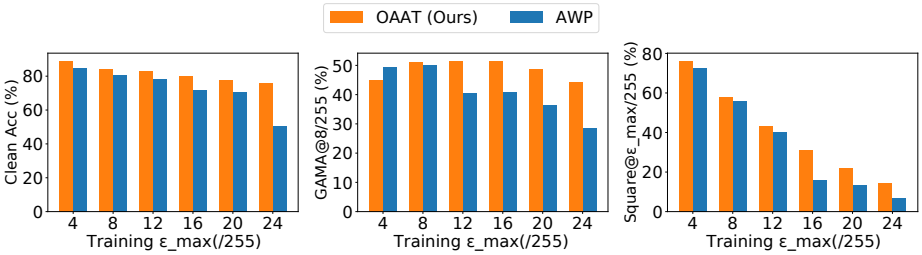

Fig. 6. **Results across variation in training ε_{max}:** While the proposed approach works best at moderate-ε bounds such as 16/255 on CIFAR-10, we observe that it outperforms the baseline for various ε_{max} values $\geq 8/255$ as well.

bound attacks at $\varepsilon = 8/255$ and 16/255 that were used for training, as well as other ℓ_p norm bound attacks and common corruptions on both datasets.

The ε_{max} used for training is a system specification, which is the perturbation bound within which the model has to be robust. Thus, to validate the efficacy of the proposed approach, we train different ResNet-18 models on CIFAR-10 using different specifications of ε_{max}. From Fig.6, we observe that for various values of training ε_{max}, the proposed approach consistently outperforms AWP [30]. Training time of OA-AT is comparable with that of AWP [30]. On CIFAR-10, OA-AT takes 7 hours 16 minutes, while AWP takes 7 hours 27 minutes for 110 epochs of training on ResNet-18 using a single V100 GPU. To ensure the absence of gradient masking in the proposed approach, we present further evaluations against diverse attacks and sanity checks in Sec.8 of the Supplementary.

Ablation Study: In order to study the impact of different components of the proposed defense, we present a detailed ablative study using ResNet-18 and WideResNet-34-10 models in Table-4. We present results on the CIFAR-10 and CIFAR-100 datasets, with E1 and F1 representing the proposed approach. First, we study the efficacy of the LPIPS metric in generating Oracle-Invariant attacks. In experiment E2, we train a model without LPIPS by setting its coefficient to zero. While the resulting model achieves a slight boost in robust accuracy at $\varepsilon = 16/255$ due to the use of stronger attacks for training, there is a considerable drop in clean accuracy, and a corresponding drop in robust accuracy at $\varepsilon = 8/255$ as well. We observe a similar trend by setting the value of α to 1 as shown in E3, and by combining E2 and E3 as shown in E4. We note that E4 is similar to standard adversarial training, where the model attempts to learn consistent predictions in the ε ball around every data sample. While this works well for large ε attacks (16/255), it leads to poor clean accuracy as shown in Table-1.

We further note that the computation of LPIPS distance using an exponential weight averaged model (E1) results in better performance as compared to using the model being trained (E5). As discussed in Sec.4, we maximize loss on $x_i + 2 \cdot \widetilde{\delta}_i$ (where $\widetilde{\delta}_i$ is the attack) in the additional weight perturbation step. We present results by using the standard ε limit for the weight perturbation step as well, in

Table 4. **CIFAR-10, CIFAR-100**: Ablation experiments on ResNet-18 architecture (E1-E7) and WideResNet-34-10 (F1-F2) architecture to highlight the importance of various aspects in the proposed defense OA-AT. Performance (%) against attacks with different ε bounds is reported.

Method	CIFAR-10				CIFAR-100			
	Clean	GAMA (8/255)	GAMA (16/255)	Square (16/255)	Clean	GAMA (8/255)	GAMA (16/255)	Square (16/255)
E1: OA-AT (Ours)	80.24	**51.40**	22.73	31.16	60.27	**26.41**	10.47	14.60
E2: LPIPS weight = 0	78.47	50.60	24.05	31.37	58.47	25.94	10.91	14.66
E3: Alpha = 1	79.29	50.60	23.65	31.23	58.84	26.15	10.97	14.89
E4: Alpha = 1, LPIPS weight = 0	77.16	50.49	**24.93**	**32.01**	57.77	25.92	**11.33**	**15.03**
E5: Using Current model (without WA) for LPIPS	**80.50**	50.75	22.90	30.76	59.54	26.23	10.50	14.86
E6: Without 2*eps perturbations for AWP	79.96	50.50	22.61	30.60	60.18	26.27	10.15	14.20
E7: Maximizing KL div in the AWP step	81.19	49.77	21.17	29.39	59.48	25.03	7.93	13.34
F1: OA-AT (Ours)	**85.32**	**58.48**	26.93	**36.93**	**65.73**	**30.90**	13.21	**18.47**
F2: LPIPS weight = 0	83.47	57.58	**27.21**	36.68	63.16	30.22	**13.59**	18.42

E6. This leads to a drop across all metrics, indicating the importance of using large magnitude perturbations in the weight perturbation step for producing a flatter loss surface that leads to better generalization to the test set. Different from the standard TRADES formulation, we maximize Cross-Entropy loss for attack generation in the proposed method. From E7 we note a drop in robust accuracy since the KL divergence based attack is weaker (Gowal et al. [11]). We present further ablative analysis in Sec.6 of the Supplementary.

8 Conclusions

In this paper, we investigate in detail robustness at larger perturbation bounds in an ℓ_p norm based threat model. We discuss the ideal goals of an adversarial defense at larger perturbation bounds, identify deficiencies of prior works in this setting and further propose a novel defense, Oracle-Aligned Adversarial Training (OA-AT) that aligns model predictions with that of an Oracle during training. The key aspects of the defense include the use of LPIPS metric for generating Oracle-Invariant attacks during training, and the use of a convex combination of clean and adversarial image predictions as targets for Oracle-Sensitive samples. We achieve state-of-the-art robustness at low and moderate perturbation bounds, and a better robustness-accuracy trade-off. We further show the relation between the contrast level of images and the existence of Oracle-Sensitive attacks within a given perturbation bound. We use this for better evaluation, and highlight the role of contrast of images in achieving an improved robustness-accuracy trade-off. We hope that future work would build on this to construct better defenses and to obtain a better understanding on the existence of adversarial examples.

9 Acknowledgements

This work was supported by a research grant (CRG/2021/005925) from SERB, DST, Govt. of India. Sravanti Addepalli is supported by Google PhD Fellowship and CII-SERB Prime Minister's Fellowship for Doctoral Research.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: The European Conference on Computer Vision (ECCV) (2020) 2, 6, 7, 10, 11
2. Balaji, Y., Goldstein, T., Hoffman, J.: Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. arXiv preprint arXiv:1910.08051 (2019) 10
3. Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A.: On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705 (2019) 1
4. Chen, J., Gu, Q.: Rays: A ray searching method for hard-label adversarial attack. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1739–1747 (2020) 6, 7, 10, 11
5. Chen, T., Zhang, Z., Liu, S., Chang, S., Wang, Z.: Robust overfitting may be mitigated by properly learned smoothening. In: International Conference on Learning Representations (ICLR) (2020) 4, 11
6. Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M.: Robustbench: a standardized adversarial robustness benchmark. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021) 11, 12
7. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning (ICML) (2020) 6, 10, 11
8. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 11
9. Goodfellow, I., Papernot, N.: Is attacking machine learning easier than defending it?, blog post on Feb 15, 2017 2, 5
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2015) 1
11. Gowal, S., Qin, C., Uesato, J., Mann, T., Kohli, P.: Uncovering the limits of adversarial training against norm-bounded adversarial examples. arXiv preprint arXiv:2010.03593 (2020) 4, 11, 14
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 11
13. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) The European Conference on Computer Vision (ECCV) (2016) 11
14. Howard, J.: Imagenette dataset (2019), https://github.com/fastai/imagenette 10
15. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407 (2018) 4, 11
16. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images (2009) 3, 10
17. Laidlaw, C., Singla, S., Feizi, S.: Perceptual adversarial robustness: Defense against unseen threat models. International Conference on Learning Representations (ICLR) (2021) 2, 7

18. Madry, A., Makelov, A., Schmidt, L., Dimitris, T., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018) 3, 4
19. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011) 10
20. Pang, T., Yang, X., Dong, Y., Su, H., Zhu, J.: Bag of tricks for adversarial training. International Conference on Learning Representations (ICLR) (2021) 4
21. Rebuffi, S.A., Gowal, S., Calian, D.A., Stimberg, F., Wiles, O., Mann, T.: Fixing data augmentation to improve adversarial robustness. arXiv preprint arXiv:2103.01946 (2021) 4, 11
22. Rice, L., Wong, E., Kolter, J.Z.: Overfitting in adversarially robust deep learning. In: International Conference on Machine Learning (ICML) (2020) 4, 11
23. Shaeiri, A., Nobahari, R., Rohban, M.H.: Towards deep learning models resistant to large perturbations. arXiv preprint arXiv:2003.13370 (2020) 4, 8
24. Sitawarin, C., Chakraborty, S., Wagner, D.: Improving adversarial robustness through progressive hardening. arXiv preprint arXiv:2003.09347 (2020) 4
25. Sriramanan, G., Addepalli, S., Baburaj, A., Venkatesh Babu, R.: Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses. In: Advances in Neural Information Processing Systems (NeurIPS) (2020) 2, 4, 6, 10, 11
26. Stutz, D., Hein, M., Schiele, B.: Relating adversarially robust generalization to flat minima. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 4
27. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR) (2013) 1, 4
28. Tramèr, F., Behrmann, J., Carlini, N., Papernot, N., Jacobsen, J.H.: Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In: International Conference on Machine Learning (ICML) (2020) 2, 5
29. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: International Conference on Learning Representations (ICLR) (2019) 6, 9
30. Wu, D., Xia, S.T., Wang, Y.: Adversarial weight perturbation helps robust generalization. Advances in Neural Information Processing Systems (NeurIPS) (2020) 3, 4, 8, 11, 12, 13
31. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 4
32. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016) 11
33. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning (ICML) (2019) 3, 4, 8
34. Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., Kankanhalli, M.: Attacks which do not kill training make adversarial learning stronger. In: International Conference on Machine Learning (ICML) (2020) 3, 4
35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 7