# Supplementary Material: 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

## 1 Oracle-Invariant Attacks

### 1.1 Square Attack

The strongest Oracle-Invariant examples are generated using the Square attack [1]. Images so generated are Oracle-Invariant since the Square Attack is query-based, and does not utilise gradients from the model for attack generation. However this attack uses 5000 queries, and is thus computationally expensive. Hence it cannot be directly incorporated for adversarial training, although it is one of the strongest attacks for evaluation purposes. We note that the computationally efficiency can be improved by reducing the number of queries; however it also reduces the effectiveness of the attack significantly. The adversarial images generated using the Square attack and their corresponding perturbations are presented in Fig.1.

### 1.2 RayS Attack

Another technique that is observed to generate strong Oracle-Invariant examples is the black-box RayS attack [4]. Similar to the Square attack, the images so generated are also Oracle-Invariant since it is a query-based attack and does not utilise gradients for attack generation. Although the RayS attack requires 10000 queries which is highly demanding from a computational viewpoint, it is observed to be weaker than the Square attack. Adversarial images generated using the RayS attack and their corresponding perturbations are presented in Fig.2.

### 1.3 PGD based Attacks

While the most efficient attack that is widely used for adversarial training is the PGD 10-step attack, it cannot be used for the generation of Oracle-Invariant samples as adversarially trained models have perceptually aligned gradients, and tend to produce Oracle-Sensitive samples. Therefore, we explore some variants

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

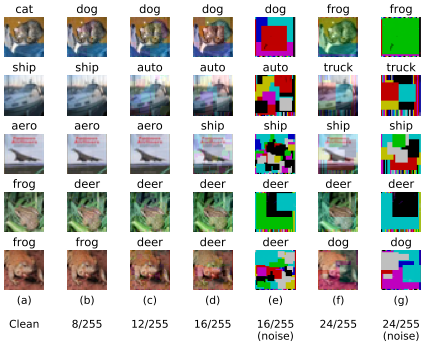


Fig. 1. **Square attack:** Adversarially attacked images (b, c, d, f) and the corresponding perturbations (e, g) for various $\ell_\infty$ bounds generated using the gradient-free random search based attack Square [1]. The clean image is shown in (a). Attacks are generated from a model trained using the proposed Oracle-Aligned Adversarial Training (OA-AT) algorithm on CIFAR-10. Prediction of the same model is printed above each image.

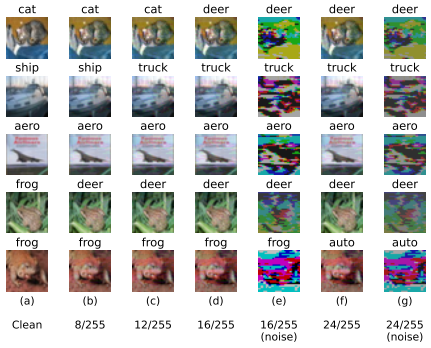


Fig. 2. **RayS attack:** Adversarially attacked images (b, c, d, f) and the corresponding perturbations (e, g) for various $\ell_\infty$ bounds generated using the gradient-free binary search based attack RayS [4]. The clean image is shown in (a). Attacks are generated from a model trained using the proposed Oracle-Aligned Adversarial Training (OA-AT) algorithm on CIFAR-10. Prediction of the same model is printed above each image.

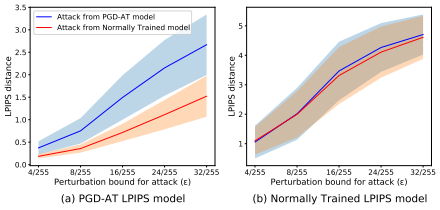


Fig. 3. LPIPS distance between clean and adversarially perturbed images. Attacks generated from PGD-AT [16,18] model (Oracle-Sensitive) and Normally Trained model (Oracle-Invariant) are considered. (a) PGD-AT ResNet-18 model is used for computation of LPIPS distance (b) Normally Trained AlexNet model is used for computation of LPIPS distance. PGD-AT model based LPIPS distance is useful to distinguish between Oracle-Sensitive and Oracle-Invariant attacks.

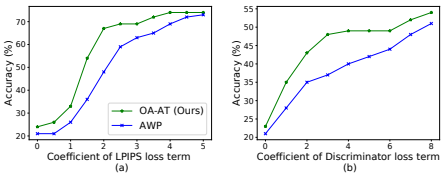


Fig. 4. Comparison of the proposed model with AWP [28] on CIFAR-10, against attacks of varying strength and Oracle sensitivity constrained within $\varepsilon = 16/255$. (a) LPIPS based regularizer, (b) Discriminator based regularizer are used for generating Oracle-Invariant attacks respectively. As the coefficient of the regularizer increases, the attack transforms from Oracle-Sensitive to Oracle-Invariant. The proposed method achieves improved accuracy when compared to AWP.

of the PGD attack to make the generated perturbations Oracle-Invariant. We denote the Cross-Entropy loss on a data sample $x$ with ground truth label $y$ using $\mathcal{L}_{CE}(x, y)$. We explore the addition of regularizers to the Cross-Entropy loss weighted by a factor of $\lambda_X$ in each case. The value of $\lambda_X$ is chosen as the minimum value which transforms the PGD attacks from Oracle-Sensitive to Oracle-Invariant. This results in the strongest possible Oracle-Invariant attacks.

### 1.4 Discriminator based PGD Attack

We train a discriminator to distinguish between Oracle-Invariant and Oracle-Sensitive adversarial examples, and further maximize the below loss for the generation of Oracle-Invariant attacks:

$$\mathcal{L}_{CE}(x, y) - \lambda_{Disc} \cdot \mathcal{L}_{BCE}(\hat{x}, \text{OI}) \tag{1}$$

Here $\mathcal{L}_{BCE}(\hat{x}, \text{OI})$ is the Binary Cross-Entropy loss of the adversarial example $\hat{x}$ w.r.t. the label corresponding to an Oracle-Invariant (OI) attack. We train the discriminator to distinguish between two input distributions; the first corresponding to images concatenated channel-wise with their respective Oracle-Sensitive perturbations, and a second distribution where perturbations are shuffled across images in the batch. This ensures that the discriminator relies on the spatial correlation between the image and its corresponding perturbation for the classification task, rather than the properties of the perturbation itself. The attack in Eq.1 therefore attempts to break the most salient property of Oracle-Sensitive attacks, which is the spatial correlation between an image and its perturbation.

### 1.5 LPIPS based PGD Attack

We propose to use the Learned Perceptual Image Patch Similarity (LPIPS) measure for the generation of Oracle-Sensitive attacks, as it is known to match well with perceptual similarity [31,15]. As shown in Fig.3, while the standard AlexNet model that is used in prior work [15] fails to distinguish between Oracle-Invariant and Oracle-Sensitive samples, an adversarially trained model is able to distinguish between the two types of attacks effectively. In this plot, we consider attacks generated from a PGD-AT [16,18] model (Fig.1(c-e) in the Main paper) as Oracle-Sensitive attacks, and attacks generated from a Normally Trained model (Fig.1(h) in the Main paper) as Oracle-Invariant attacks. We therefore propose to minimize the LPIPS distance between the natural and perturbed images, in addition to the maximization of Cross-Entropy loss for attack generation as shown below:

$$\mathcal{L}_{CE}(x, y) - \lambda_{\text{LPIPS}} \cdot \text{LPIPS}(x, \hat{x}) \tag{2}$$

We choose $\lambda_{\text{LPIPS}}$ as the minimum value that transforms the PGD attack from Oracle-Sensitive to Oracle-Invariant (OI), to generate strong OI attacks. This is further fine-tuned during training to achieve the optimal robustness-accuracy trade-off. As shown in Fig.5, setting $\lambda_{\text{LPIPS}}$ to 1 changes adversarial

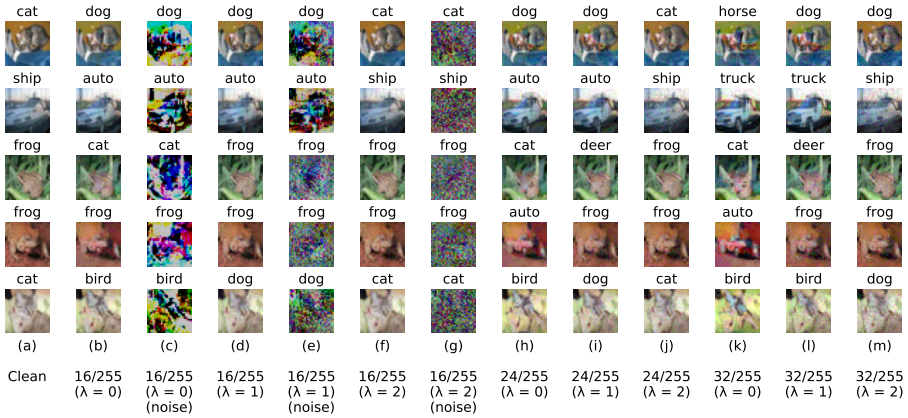


Fig. 5. Oracle-Invariant adversarial examples generated using the LPIPS based PGD attack in Eq.2 across various perturbation bounds. White-box attacks and predictions on the model trained using the proposed OA-AT defense on the CIFAR-10 dataset with ResNet-18 architecture are shown: (a) Original Unperturbed image, (b, h, k) Adversarial examples generated using the standard PGD 10-step attack, (d, f, i, j, l, m) LPIPS based PGD attack generated within perturbation bounds of 16/255 (d, f), 24/255 (i, j) and 32/255 (l, m) by setting the value of $\lambda_{\text{LPIPS}}$ to 1 and 2, (c, e, g) Perturbations corresponding to (b), (d) and (f) respectively.

examples from Oracle-Sensitive to Oracle-Invariant, as they look similar to the corresponding original images shown in Fig.5 (a). This can be observed more distinctly at perturbation bounds of 24/255 and 32/255. The perturbations in Fig.5 (c) are smooth, while those in (e) and (g) are not. This shows that the addition of the LPIPS term helps in making the perturbations Oracle-Invariant. Very large coefficients of the LPIPS term make the attack weak as can be seen in Fig.5 (f, j, m) where the model prediction is same as the true label. We therefore set $\lambda_{\text{LPIPS}}$ to 1 to obtain strong Oracle-Invariant attacks. As shown in Table-1, while we obtain the best results using the LPIPS based PGD attack for training (E1), the use of discriminator based PGD attack (E8) also results in a better robustness-accuracy trade-off when compared to E2, where there is no explicit regularizer to ensure the generation of Oracle-Invariant attacks.

### 1.6 Evaluation of the proposed defense against Oracle-Invariant Attacks

We compare the performance of the proposed defense OA-AT with the strongest baseline AWP [28] against the two proposed Oracle-Invariant attacks, LPIPS based attack and Discriminator based attack in Fig.4 (a) and (b) respectively. We vary the coefficient of the regularizers used in the generation of attacks, $\lambda_{Disc}$ (Eq.1) and $\lambda_{\text{LPIPS}}$ (Eq.2) in each of the plots. As we increase the coefficient, the attack transforms from Oracle-Sensitive to Oracle-Invariant. The

proposed method (OA-AT) achieves improved accuracy compared to the AWP [28] baseline.

## 2 Analysing Oracle Alignment of Adversarial Attacks

In this section, we present more detailed analysis of generating Oracle-Invariant and Oracle-Sensitive attacks in a simplified yet natural setting, introduced in Sec.5 of the Main paper. We consider a binary classification task as proposed by Tsipras et al. [26], 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.

Case 1: A simple, yet effective classifier that achieves high accuracy is given as follows: $f(x) := sign(w^T x)$, where $w \in \mathbb{R}^{d+1}$ with $w = (0, \frac{1}{d}, \ldots, \frac{1}{d})$. Then, its accuracy is given by

$$\mathbb{P}[f(x) = y] = \mathbb{P}[y \cdot sign(w^T x) = 1] = \mathbb{P}\left[y \sum_{i=2}^{d+1} \frac{x_i}{d} > 0\right] = \mathbb{P}[z > 0]$$

where $z \sim \mathcal{N}(\alpha, \frac{1}{d})$.
We see this is true since $z$ is given by a sum of $d$ i.i.d. Gaussian random variables $yx_i/d$, each with mean $y \cdot \alpha y/d = \alpha/d$ (since $y^2 = 1$), and with variance $1/d^2$ each. Thus the accuracy exceeds 99% if $\alpha > \frac{3}{\sqrt{d}}$ by properties of the Gaussian distribution. We note that such a classifier that achieves vanishing error can be learnt through standard Empirical Risk Minimisation.

However, with an $\ell_\infty$ attack with perturbation budget $2\alpha$, an adversary can flip each of the weakly correlated features $x_2, \ldots, x_{d+1}$ to appear as $x_i + \delta \sim \mathcal{N}(-\alpha y, 1)$. These perturbed features are thus now weakly anti-correlated with the Oracle label $y$, and achieves a robust accuracy of less than 1%. Thus by attacking a standard, non-robust classifier the perturbations can be seen to be Oracle Invariant features that are spurious in nature. Tsipras et al. [26] prove a more general version of the same result:

**Theorem [Tsipras et al.]** Let $f$ be any classifier that achieves standard accuracy of at least $1 - \gamma$, that is $\mathbb{P}[f(x) = y] > 1 - \gamma$. Then, the robust accuracy achieved by $f$ under an $\ell_\infty$ attack with a perturbation budget of $2\alpha$, has a tight upper-bound given by $\frac{p}{1-p}\gamma$.

**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 $y$.

Case 2: Here, we consider another simple classifier, $g(x) := sign(x_1)$, which achieves, on natural samples, an accuracy of

$$\mathbb{P}[g(x) = y] = \mathbb{P}[y \cdot sign(x_1) = 1] = p$$

While the accuracy thus has a tight upper-bound of $p$, the classifier $g$ is robust to adversarial perturbations of relatively large magnitude. Further, to maximise the misclassification of model $g$, it is easy to see that adversarial perturbations take the form $\delta = (2\alpha, 0, \ldots, 0)$. Thus, we observe that adversarial perturbations of robust models correspond to Oracle Sensitive features.

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

**Theorem 1.** *Consider a robust Deep Neural Network $f_\theta$ with parameters $\theta$ as in Algorithm-1. Given an input sample $x$, let $\delta^*$ represent an optimal solution that maximises the following objective:*

$$\ell = \ell_{CE}(f_\theta(x + \delta), y) - \lambda \cdot \text{LPIPS}(x, x + \delta) \tag{3}$$

*Then, $\exists\ \lambda > 0$ such that $x + \delta^*$ is an Oracle Invariant Sample.*

**Proof:** By definition, the LPIPS metric between samples $x$ and $x + \delta^*$ measures aggregate L2 distances between the corresponding feature space representations in the intermediate layers of robust network $f_\theta$. For $\lambda >> 0$, the LPIPS component dominates the overall optimization objective in the adversarial attack. To prove the result, let us assume on the contrary that the perturbation $\delta^*$ results in an Oracle Sensitive Sample. Thus the corresponding feature representations in a robust network for the sample $x + \delta^*$ would deviate significantly from that of the original benign sample $x$. Thus, as $\text{LPIPS}(x, x + \delta^*) > 0$, and as $\lambda \to \infty$, the overall objective in Eqn(1) decreases, with $\ell \to -\infty$, contradicting the optimality of $\delta^*$ in maximising the same objective. Thus, we conclude that $x + \delta^*$ is indeed an Oracle Invariant Sample.

## 3 Details on the Datasets used

We evaluate the proposed approach on the CIFAR-10, CIFAR-100 [14] and SVHN [17] datasets. The three datasets consist of RGB images of spatial dimension 32×32. CIFAR-10 and SVHN contain 10 distinct classes, while CIFAR-100 contains 100. CIFAR-10 is the most widely used benchmark dataset to perform a comparative analysis across different adversarial defense and attack methods. CIFAR-100 is a challenging dataset to achieve adversarial robustness given the large number of diverse classes that are interrelated. Each of these datasets consists of 50,000 training images and 10,000 test images, while SVHN contains

73257 training and 26032 testing images. We split the original training set to create a validation set of 1,000 images in CIFAR-10 and 2,500 images in CIFAR-100 and SVHN. We ensure that the validation split is balanced equally across all classes, and use the remaining images for training. To ensure a fair comparison, we use the same split for training the proposed defense as well as other baseline approaches. For CIFAR-10 and CIFAR-100 datasets, we consider the $\ell_\infty$ threat model of radius 8/255 to be representative of imperceptible perturbations, that is, the Oracle label does not change within this set. For SVHN we consider this bound to be 4/255 as many of the images in the dataset have a low contrast, leading to visible perturbations at relatively small $\varepsilon$ bounds. Further, to investigate robustness within moderate magnitude perturbation bounds, we consider the $\ell_\infty$ threat model of radius 16/255 for CIFAR-10 and CIFAR-100, and a bound of 12/255 for SVHN.

## 4 Details on Training

In this section, we expound further details on the algorithm of the proposed method, presented in Sec.4 of the Main paper (Alg.1). We use a varying $\varepsilon$ schedule and start training on perturbations of magnitude $\varepsilon_{max}/4$. This results in marginally better performance when compared to ramping up the value of $\varepsilon$ from 0 (E9 of Table-1). For CIFAR-10 training on ResNet-18, we set the weight of the adversarial loss $L_{adv}$ in L21 of Alg.1 ($\beta$ parameter of TRADES [29]) to 1.5 for the first three-quarters of training, and then linearly increase it from 1.5 to 3 in the moderate perturbation regime, where $\varepsilon$ is linearly increased from 12/255 to 16/255. In this moderate perturbation regime, we also linearly increase the coefficient of the LPIPS distance (Alg.1, L14) from 0 to 1, and linearly decrease the $\alpha$ parameter used in the convex combination of softmax prediction (Alg.1, L11) from 1 to 0.8. This results in a smooth transition from adversarial training on imperceptible attacks to attacks with larger perturbation bounds. We set the weight decay to 5e-4.

We use cosine learning rate schedule with a maximum learning rate of 0.2 for CIFAR-10 and CIFAR-100, and 0.05 for SVHN. We use SGD optimizer with momentum of 0.9, and train for 110 epochs, except for training PreActResNet18 on CIFAR-100 and WideResNet-34-10 on CIFAR-10, where we use 200 epochs. We do not perform early stopping and always report accuracy of the last epoch. We compute the LPIPS distance using an exponential weight averaged model with $\tau = 0.995$. We note from Table-1 that the use of weight-averaged model (E1) results in better performance when compared to using the model being trained for the same (E5). This also leads to more stable results across reruns.

We utilise AutoAugment [8] for training on CIFAR-100, SVHN and for CIFAR-10 training on large model capacities. We apply AutoAugment with a probability of 0.5 for CIFAR-100, and for the CIFAR-10 model trained on ResNet-34. Since the extent of overfitting is higher for large model capacities, we use AutoAugment with $p = 1$ on WideResNet-34-10. While the use of AutoAugment helps in overcoming overfitting, it could also negatively impact ro-

Table 1. **CIFAR-10, CIFAR-100**: Ablation experiments on ResNet-18 architecture to highlight the importance of various aspects in the proposed defense OA-AT. Performance (%) against attacks with different perturbation bounds $\varepsilon$ 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 |
| **E8**: Using Discriminator instead of LPIPS (OI Attack) | 80.56 | 50.75 | 22.13 | 31.17 | 58.84 | 26.35 | 10.64 | 14.82 |
| **E9**: Increasing epsilon from the beginning | 80.34 | 50.77 | 22.57 | 30.80 | **60.51** | 26.34 | 10.37 | 14.61 |
| **E10**: Without AutoAugment | 80.24 | **51.40** | 22.73 | 31.16 | 58.08 | 25.81 | 10.40 | 14.31 |
| **E11**: With AutoAugment (p=0.5) | 81.59 | 50.40 | 21.59 | 30.84 | 60.27 | **26.41** | 10.47 | 14.60 |
| **E12**: With AutoAugment (p=1) | **81.74** | 48.15 | 18.92 | 28.31 | 60.19 | 25.32 | 9.24 | 13.78 |
| **E13**: Alpha = 1, LPIPS weight = 0 + fixed $\varepsilon$=16/255 | 71.64 | 47.59 | 25.91 | 31.75 | 50.99 | 23.19 | 9.99 | 13.48 |

bust accuracy due to the drift between the training and test distributions. We observe a drop in robust accuracy on the CIFAR-10 dataset with the use of AutoAugment (E11, E12 in Table-1), while there is a boost in the clean accuracy. On similar lines, we observe a drop in robust accuracy on the CIFAR-100 dataset as well, when we increase the probability of applying AutoAugment from 0.5 (E11 in Table-1) to 1 (E12 in Table-1). We use AutoAugment with $p = 1$ for SVHN, as we observe that it results in more stable training. Further, we find that using Label Smoothing with CIFAR-100 helps in improving clean accuracy, as shown in Table-2 of the Main paper.

To investigate the stability of the proposed approach, we train a ResNet-18 network multiple times by using different random initialization of network parameters. We observe that the proposed approach is indeed stable, with standard deviation of 0.167, 0.115, 0.180 and 0.143 for clean accuracy, GAMA PGD-100 accuracies with $\varepsilon = 8/255$ and $16/255$, and accuracy against the Square attack with $\varepsilon = 16/255$ respectively over three independent training runs on CIFAR-10. We also observe that the last epoch is consistently the best performing model for the ResNet-18 architecture. Nonetheless, we still utilise early stopping on the validation set using PGD 7-step accuracy for all the baselines to enable a fair comparison overall.

## 5 Evaluation of Adversarial Defenses

Gradient-based white-box attacks such as PGD [16], GAMA-PGD [22] and Auto-PGD with Cross-Entropy (CE) and Difference of Logits Ratio (DLR) losses [6] are known to be the strongest attacks against standard Adversarial defenses that do not obfuscate gradients. Gradient-Free attacks such as ZOO [5], SPSA [27], Square [1] and RayS [4] are useful to craft perturbations without requiring white-box access to the model. These attacks are also used to reliably estimate the robustness of defenses that rely on gradient masking [19]. Amongst

the Gradient-Free attacks, Square and Ray-S do not use Zeroth order gradient estimates, and utilize Random-Search and Binary-Search based algorithms respectively to construct strong attacks against a given defense. We use such query-based attacks to generate perturbations that do not flip Oracle predictions even for moderate-magnitude constraint sets. AutoAttack combines strong untargeted and targeted white-box attacks with the query-based black-box attack Square to effectively estimate the robustness of a given defense, and is a well accepted standard for benchmarking defenses. We report our results against GAMA-PGD, AutoAttack, Square and Ray-S. We also present further evaluations using various adaptive attacks (Sec.6 of the Supplementary) to reliably estimate robustness of the proposed defense.

## 6 Ablation Study

In order to study the impact of different components of the proposed defense, we present a detailed ablative study using ResNet-18 models in Table-1. We present results on the CIFAR-10 and CIFAR-100 datasets, with E1 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 $\alpha$ 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 $\varepsilon$ ball around every data sample. While this works well for large $\varepsilon$ attacks ($\varepsilon = 16/255$), it leads to poor clean accuracy as shown in the first partition of Table-2.

As discussed in Sec.4 of the Main paper, 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 $\varepsilon$ limit for the weight perturbation step as well, in 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 that the use of KL divergence leads to a drop in robust accuracy since the KL divergence based attack is weaker. This is consistent with the observation by Gowal et al. [10]. However, on the SVHN dataset, we find that the use of KL divergence based attack results in a significant improvement in clean accuracy, leading to better robust accuracy as well. We therefore utilize the KL divergence loss for attack generation on the SVHN dataset. We also investigate the effect of AutoAugment [8], Weight Averaging [12] and Label Smoothing + Warmup on the AWP [28] baseline in Table-6.

Table 2. **CIFAR-10**: Performance (%) of the proposed defense OA-AT against attacks bounded within different $\varepsilon$ bounds, when compared to the following baselines: AWP [28], ExAT [20], TRADES [29], ATES [21], PGD-AT [16] and FAT [30]. AWP [28] is the strongest baseline. The first partition shows defenses trained on $\varepsilon = 16/255$. Training on large perturbation bounds results in very poor Clean Accuracy. The second partition consists of baselines tuned to achieve clean accuracy close to 80%. These are sorted by AutoAttack accuracy [7] (AA 8/255). The proposed defense (OA-AT) achieves significant gains in accuracy across all attacks.

| Method | Attack $\varepsilon$ (Training) | Clean | FGSM (BB) (8/255) | R-FGSM (8/255) | GAMA (8/255) | AA (8/255) | FGSM (BB) (12/255) | R-FGSM (12/255) | GAMA (12/255) | Square (12/255) | FGSM (BB) (16/255) | R-FGSM (16/255) | GAMA (16/255) | Square (16/255) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TRADES | 16/255 | 75.30 | 73.26 | 53.10 | 35.64 | 35.12 | 72.13 | 44.27 | 20.24 | 30.11 | 70.76 | 36.99 | 10.10 | 18.87 |
| AWP | 16/255 | 71.63 | 69.71 | 54.53 | 40.85 | 40.55 | 68.65 | 47.13 | 27.06 | 34.42 | 67.42 | 40.89 | 15.92 | 24.16 |
| PGD-AT | 16/255 | 64.93 | 63.65 | 55.47 | 46.66 | 46.21 | 62.81 | 51.05 | 36.95 | 40.53 | 61.70 | 46.40 | 26.73 | 32.25 |
| FAT | 16/255 | 75.27 | 73.44 | 60.25 | 47.68 | 47.34 | 72.22 | 53.17 | 34.31 | 39.79 | 70.73 | 46.88 | 22.93 | 29.47 |
| ExAT+AWP | 16/255 | 75.28 | 73.27 | 60.02 | 47.63 | 47.46 | 71.81 | 52.38 | 34.42 | 39.62 | 70.47 | 45.39 | 22.61 | 28.79 |
| ATES | 16/255 | 66.78 | 65.60 | 56.79 | 47.89 | 47.52 | 64.64 | 51.71 | 37.47 | 42.07 | 63.75 | 47.28 | 26.50 | 32.55 |
| ExAT + PGD | 16/255 | 72.04 | 70.68 | 59.99 | 49.24 | 48.80 | 69.66 | 53.96 | 36.68 | 41.93 | 68.04 | 48.37 | 23.01 | 30.21 |
| FAT | 12/255 | 80.27 | 77.87 | 61.46 | 45.42 | 45.13 | 76.69 | 52.33 | 29.08 | 36.71 | 74.79 | 44.56 | 16.18 | 24.59 |
| FAT | 8/255 | **84.36** | 82.20 | 64.06 | 48.41 | 48.14 | 80.32 | 55.41 | 29.39 | 39.48 | 78.13 | 47.50 | 15.18 | 25.07 |
| ATES | 8/255 | 84.29 | **82.39** | **65.66** | 49.14 | 48.56 | **80.81** | 55.59 | 29.36 | 40.68 | **78.48** | 47.03 | 14.70 | 25.88 |
| PGD-AT | 8/255 | 81.12 | 78.94 | 63.48 | 49.03 | 48.58 | 77.19 | 54.42 | 30.84 | 40.82 | 74.37 | 46.28 | 15.77 | 26.47 |
| PGD-AT | 10/255 | 79.38 | 77.89 | 62.78 | 49.28 | 48.68 | 76.60 | 54.76 | 32.40 | 41.46 | 74.75 | 47.46 | 18.18 | 28.29 |
| AWP | 10/255 | 80.32 | 77.87 | 62.33 | 49.06 | 48.89 | 76.33 | 53.83 | 32.88 | 40.27 | 74.13 | 45.51 | 19.17 | 27.56 |
| ATES | 10/255 | 80.95 | 79.22 | 63.95 | 49.57 | 49.12 | 77.77 | 55.37 | 32.44 | 42.21 | 75.51 | 48.12 | 18.36 | 29.07 |
| TRADES | 8/255 | 80.53 | 78.58 | 63.69 | 49.63 | 49.42 | 77.20 | 55.48 | 33.32 | 40.94 | 75.05 | 47.92 | 19.27 | 27.82 |
| ExAT + PGD | 11/255 | 80.68 | 79.07 | 63.58 | 50.06 | 49.52 | 77.98 | 55.92 | 32.47 | 41.10 | 76.12 | 48.37 | 17.81 | 27.23 |
| ExAT + AWP | 10/255 | 80.18 | 78.04 | 63.15 | 49.87 | 49.69 | 76.34 | 54.64 | 33.51 | 41.04 | 74.37 | 46.54 | 20.04 | 28.40 |
| AWP | 8/255 | 80.47 | 78.22 | 63.32 | 50.06 | 49.87 | 76.88 | 54.61 | 33.47 | 41.05 | 74.42 | 46.16 | 19.66 | 28.51 |
| **OA-AT (Ours)** | 16/255 | 80.24 | 78.54 | 65.00 | **51.40** | **50.88** | 77.34 | **57.68** | **36.01** | **43.20** | 75.72 | **51.13** | **22.73** | **31.16** |
| Gain w.r.t. AWP | | −0.23 | +0.32 | +1.68 | +1.34 | +1.01 | +0.46 | +3.07 | +2.54 | +2.15 | +1.30 | +4.97 | +3.07 | +2.65 |

## 7 Detailed Results

In Tables-2 and 3, we present results of different defense methods such as AWP-TRADES [28], TRADES [29], PGD-AT [16], ExAT [20], ATES [21] and FAT [30], evaluated across a wide range of adversarial attacks. We present evaluations on the Black-Box FGSM attack [9] and a suite of White-Box attacks, on $\ell_\infty$ constraint sets of different radii: 8/255, 12/255 and 16/255. The white-box evaluations consist of the single-step Randomized-FGSM (R-FGSM) attack [25], the GAMA PGD-100 attack [22] and AutoAttack [7], with the latter two being amongst the strongest of attacks known to date. Lastly, we also present evaluations on the Square attack [1] for $\varepsilon = 12/255$ and 16/255 in order to evaluate performance on Oracle-Invariant samples at large perturbation bounds.

### 7.1 CIFAR-10

To enable a fair comparison of the proposed approach with existing methods, we present comprehensive results of various defenses trained with different attack strengths in Table-2. In the first partition of the table, we present baselines trained using attacks constrained within an $\ell_\infty$ bound of 16/255. While these models do achieve competitive robustness on adversaries of attack strength $\varepsilon = 8/255$, 12/255 and 16/255, they achieve significantly lower accuracy on clean samples which limits their use in practical scenarios. Thus, for better comparative analysis that accounts for the robustness-accuracy trade-off, we present results of the existing methods with hyperparameters and attack strengths tuned

Table 3. **CIFAR-100**: Performance (%) of the proposed defense OA-AT against attacks bounded within different $\varepsilon$ bounds, when compared to the following baselines: AWP [28], ExAT [20], TRADES [29], ATES [21], PGD-AT [16] and FAT [30]. AWP [28] is the strongest baseline. The baselines are sorted by AutoAttack accuracy [7] (AA 8/255). The proposed defense achieves significant gains in accuracy against the strongest attacks across all $\varepsilon$ bounds. Since the proposed defense uses AutoAugment [8] as the augmentation strategy, we present results on the strongest baseline AWP [28] with AutoAugment as well.

| Method | Attack $\varepsilon$ (Train) | Clean | FGSM-BB (8/255) | R-FGSM (8/255) | GAMA (8/255) | AA (8/255) | FGSM-BB (12/255) | R-FGSM (12/255) | GAMA (12/255) | Square (12/255) | FGSM-BB (16/255) | R-FGSM (16/255) | GAMA (16/255) | Square (16/255) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FAT | 8/255 | 56.61 | 52.10 | 34.76 | 23.36 | 23.20 | 49.54 | 27.77 | 13.96 | 18.21 | 46.01 | 22.52 | 8.30 | 11.56 |
| TRADES | 8/255 | 58.27 | 54.33 | 36.20 | 23.67 | 23.47 | 51.64 | 28.55 | 13.88 | 18.46 | 48.46 | 22.78 | 8.31 | 11.89 |
| PGD-AT | 8/255 | 57.43 | 53.71 | 37.66 | 24.81 | 24.33 | 50.90 | 30.07 | 13.51 | 19.62 | 47.43 | 23.18 | 7.40 | 11.64 |
| ATES | 8/255 | 57.54 | 53.62 | 37.05 | 25.08 | 24.72 | 50.84 | 29.18 | 13.75 | 19.42 | 47.35 | 22.89 | 7.59 | 11.40 |
| ExAT-PGD | 9/255 | 57.46 | 53.56 | 38.48 | 25.25 | 24.93 | 51.43 | 30.60 | 15.12 | 20.40 | 48.15 | 24.21 | 8.37 | 12.47 |
| ExAT-AWP | 10/255 | 57.76 | 53.46 | 37.84 | 25.55 | 25.27 | 50.42 | 30.39 | 14.98 | 19.72 | 46.99 | 24.48 | 9.07 | 12.68 |
| AWP | 8/255 | 58.81 | 54.13 | 37.92 | 25.51 | 25.30 | 50.72 | 30.40 | 14.71 | 19.82 | 46.66 | 23.96 | 8.68 | 12.44 |
| AWP (with AutoAug.) | 8/255 | 59.88 | 55.62 | 39.10 | 25.81 | 25.52 | 52.75 | 31.11 | 14.80 | 20.24 | 49.44 | 24.99 | 8.72 | 12.80 |
| **Ours (with AutoAug)** | 16/255 | **60.27** | **56.27** | **40.24** | **26.41** | **26.00** | **53.86** | **33.78** | **16.28** | **21.47** | **51.11** | **28.02** | **10.47** | **14.60** |
| Gain w.r.t. AWP (with AutoAug) | | +0.39 | +0.65 | +1.14 | +0.60 | +0.48 | +1.11 | +2.67 | +1.48 | +1.23 | +1.67 | +3.03 | +1.75 | +1.80 |

to achieve the best robust performance, while maintaining clean accuracy close to 80% as commonly observed on the CIFAR-10 dataset on ResNet-18 architecture, in the second partition of Table-2. We observe that the proposed method OA-AT consistently outperforms other approaches on all three metrics described in Sec.3.3 of the Main paper, by achieving enhanced performance at $\varepsilon = 8/255$ and 16/255, while striking a favourable robustness-accuracy trade-off as well. The proposed defense achieves better robust performance even on the standard $\ell_\infty$ constraint set of 8/255 when compared to existing approaches, despite being trained on larger perturbations sets.

### 7.2 CIFAR-100

In Table-3, we present results on models trained on the highly-challenging CIFAR-100 dataset. Since this dataset contains relatively fewer training images per class, we seek to enhance performance further by incorporating the augmentation technique, AutoAugment [8,23]. To enable fair comparison, we incorporate AutoAugment for the strongest baseline, AWP [28] as well. We observe that the proposed method consistently performs better than existing approaches by significant margins, both in terms of clean accuracy, as well as robustness against adversarial attacks conforming to the three distinct constraint sets. Further, this also confirms that the proposed method scales well to large, complex datasets, while maintaining a consistent advantage in performance compared to other approaches.

### 7.3 $\ell_2$ Threat Model

OA-AT indeed works well when trained on $\ell_2$ adversaries, as shown in Table-4. While the standard perturbation bound considered for $\ell_2$ norm is 0.5, we show significant improvements for $\varepsilon$= 0.75, 1 as well. We obtain consistent gains at

large $\varepsilon$ bounds while achieving similar clean accuracy and robust accuracy at the lower bound of 0.5. On CIFAR-100, we obtain 2.5%-3% gains across perturbation bounds of 0.5, 0.75 and 1.

### 7.4 AWP+ results

Since the proposed method utilizes additional techniques to overcome overfitting and improve generalization, we generate improved baselines as well, using the same techniques, in order to facilitate a fair comparison. In Table-6, we present results of improved AWP [28] baselines by using AutoAugment [8,23], Weight Averaging [12] and Label Smoothing + Warmup. As previously seen from evaluations as reported in Table-2 of the Main paper, we again observe that the proposed method OA-AT consistently outperforms all the AWP+ baselines presented in Table-6. Further, we present OA-AT and AWP [28] with and without Weight Averaging [12] in Table-5, by training WideResNet-34-10 models. We observe that Weight Averaging does not lead to significant gains for both OA-AT as well as AWP [28].

## 8 Gradient Masking Checks

As discussed by Athalye et al. [2], we present various checks to ensure the absence of Gradient Masking in the proposed defense. In Fig.6 (a,c), we observe that the accuracy of the proposed defense on the CIFAR-10 and CIFAR-100 datasets monotonically decreases to zero against 7-step PGD white-box attacks as the perturbation budget is increased. This shows that gradient based attacks indeed serve as a good indicator of robust performance, as strong adversaries of large perturbation sizes achieve zero accuracy, indicating the absence of gradient masking. In Fig.6 (b,d), we plot the Cross-Entropy loss against FGSM attacks with varying perturbation budget. We observe that the loss increases linearly, thereby suggesting that the first-order Taylor approximation to the loss surface indeed remains effective in the local neighbourhood of sample images, again indicating the absence of gradient masking.

We verify that the model achieves higher robust accuracy against weaker Black-box attacks, when compared to strong gradient based attacks such as GAMA or AutoAttack in Tables-2, 3. We also observe that adversaries that conform to larger constraint sets are stronger than their counterparts that are restricted to smaller epsilon bounds, as expected.

Table 4. Prediction accuracy (%) of PreActResNet-18 models trained using TRADES-AWP and OA-AT on $\ell_2$ adversaries.

| Dataset | Method | Clean | AA@0.5 | Square@0.75 | Square@1 |
|---|---|---|---|---|---|
| **CIFAR-10** | TRADES-AWP | 88.45 | 71.34 | 71.19 | 64.21 |
| | OA-AT (Ours) | **89.13** | **71.40** | **73.33** | **66.48** |
| **CIFAR-100** | TRADES-AWP | 70.38 | 41.96 | 46.77 | 38.62 |
| | OA-AT (Ours) | **70.41** | **44.70** | **49.34** | **41.58** |

Table 5. **Effect of Weight Averaging [12] on AWP [28] and OA-AT**: Performance (%) of WideResNet-34-10 models trained using the proposed defense OA-AT and AWP [28], against GAMA-PGD100 [22] and Square [1] attacks with and without using Weight Averaging [12].

| Method | CIFAR-10, WRN-34-10 | | | CIFAR-100, WRN-34-10 | | |
|---|---|---|---|---|---|---|
| | Clean Acc | GAMA (8/255) | Square (16/255) | Clean Acc | GAMA (8/255) | Square (16/255) |
| AWP (without WA) | 85.36 | 56.34 | 31.70 | 62.78 | 29.82 | 15.70 |
| AWP+ (with WA) | 85.52 | 56.42 | 32.41 | 62.73 | 29.92 | 15.85 |
| OAAT (without WA) | 85.28 | 58.19 | 36.75 | 65.53 | 30.59 | 18.06 |
| Ours (OAAT with WA) | 85.32 | 58.48 | 36.93 | 65.73 | 30.90 | 18.47 |

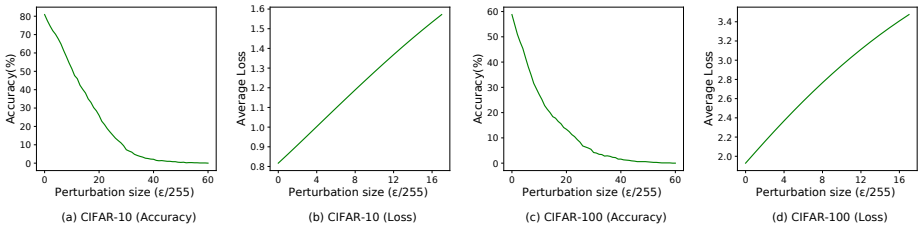


Fig. 6. Accuracy and Loss plots on a 1000-sample class-balanced subset of the respective test-sets of CIFAR-10 and CIFAR-100 datasets. (a, c) Plots showing the trend of Accuracy (%) against PGD-7 step attacks across variation in attack perturbation bound ($\varepsilon$) on CIFAR-10 and CIFAR-100 datasets with ResNet-18 architecture. As the perturbation bound increases, accuracy against white-box attacks goes to 0, indicating the absence of gradient masking [2] (b, d) Plots showing the variation of Cross-Entropy Loss on FGSM attack [9] against variation in the attack perturbation bound ($\varepsilon$). As the perturbation bound increases, loss increases linearly, indicating the absence of gradient masking [2]

In Table-7, we perform exhaustive evaluations using various attack techniques to further verify the absence of gradient masking. In addition to AutoAttack [7] which in itself consists of an ensemble of four attacks (AutoPGD with Cross-Entropy and Difference-of-Logits loss, the FAB attack [6] and Square Attack [1]), we present evaluations against strong multi-targeted attacks such as GAMA-MT [22] and the MDMT attack [13] which specifically target other classes during optimization. We also consider the untargeted versions of the latter two attacks, the GAMA-PGD and MD attack respectively. We also present robustness against the ODS attack [24] with 100 restarts, which diversifies the input random noise based on the output predictions in order to obtain results which are less dependent on the sampled random noise used for attack initialization. Next, the Logit-Scaling attack [3,11] helps yield robust evaluations that are less dependent on the exact scale of output logits predicted by the network, and is seen to be

Table 6. **Improvements to the AWP baseline:** Performance (%) of models trained by applying AutoAugment [8], Label Smoothing + Warmup and Weight Averaging [12] to the AWP baseline [28], against GAMA-PGD100 [22] and Square [1] attacks. Results on the CIFAR-10, CIFAR-100 and SVHN datasets are reported using different $\varepsilon$ bounds.

| **Auto-Augment probability** | **Label Smoothing + Warmup** | **Weight Averaging** | **Metrics of interest** | | | **Others** |
|---|---|---|---|---|---|---|
| | | | **Clean** | **GAMA (8/255)** | **Square (16/255)** | **GAMA (16/255)** |
| **CIFAR-10 (WRN-34-10), 200 epochs** | | | | | | |
| 0 | × | × | 85.36 | 56.34 | 31.54 | 23.74 |
| 0 | × | ✓ | 85.52 | 56.42 | 32.41 | 24.04 |
| 1 | × | × | 87.36 | 52.62 | 29.83 | 19.39 |
| 1 | × | ✓ | 86.75 | 53.62 | 30.11 | 20.41 |
| **CIFAR-100 (ResNet-18), 110 epochs** | | | | | | |
| 0 | × | × | 58.81 | 25.51 | 12.44 | 8.68 |
| 0.5 | × | × | 59.88 | 25.81 | 12.80 | 8.72 |
| 0 | ✓ | ✓ | 58.99 | 26.07 | 13.10 | 8.98 |
| 0.5 | ✓ | × | 59.82 | 25.39 | 13.04 | 8.62 |
| **CIFAR-100 (PreActResNet-18), 200 epochs** | | | | | | |
| 0 | × | × | 58.85 | 25.58 | 12.39 | 9.01 |
| 0.5 | ✓ | × | 62.10 | 25.99 | 13.27 | 8.91 |
| 0.5 | ✓ | ✓ | 62.11 | 26.21 | 13.26 | 9.21 |
| 0 | ✓ | × | 59.70 | 26.61 | 13.80 | 9.70 |
| 0 | ✓ | ✓ | 59.97 | 26.90 | 13.74 | 9.95 |
| **CIFAR-100 (WRN-34-10), 110 epochs** | | | | | | |
| 0 | × | × | 62.41 | 28.98 | 14.68 | 10.98 |
| 0 | × | ✓ | 61.72 | 29.78 | 15.32 | 11.15 |
| 0.5 | × | × | 61.33 | 29.22 | 15.18 | 10.94 |
| 0 | ✓ | × | 62.78 | 29.82 | 15.70 | 11.45 |
| 0 | ✓ | ✓ | 62.73 | 29.92 | 15.85 | 11.55 |
| 0.5 | ✓ | ✓ | 62.23 | 29.36 | 15.47 | 11.20 |
| **SVHN (PreActResNet-18), 110 epochs** | | | | | | |
| 0 | × | × | 91.91 | 75.92 | 35.78 | 30.70 |
| 0.5 | × | × | 90.99 | 75.37 | 36.42 | 31.02 |
| 0.5 | × | ✓ | 92.21 | 72.31 | 36.02 | 30.80 |
| 1 | × | × | 89.97 | 75.08 | 38.47 | 31.34 |
| 1 | × | ✓ | 89.71 | 74.73 | 38.41 | 31.15 |

effective on some defenses which exhibit gradient masking. However, we observe that the proposed method is robust against all such attacks, with the lowest accuracy being attained on the AutoAttack ensemble.

Table 7. **Evaluation against various attacks constrained within a perturbation bound of $\varepsilon = 8/255$ on CIFAR-10:** Performance (%) of the proposed defense OA-AT on ResNet-18 architecture against various attacks (sorted by Robust Accuracy) to ensure the absence of gradient masking.
†Includes 5000-queries of Square attack.

| Attack | No. of Steps | No. of restarts | Robust Accuracy (%) |
|---|---|---|---|
| AutoAttack† [7] | 100 | 20 | **50.88** |
| GAMA-MT [22] | 100 | 5 | 50.90 |
| ODS (98 +2 steps) [24] | 100 | 100 | 50.94 |
| MDMT attack [13] | 100 | 10 | 51.19 |
| Logit-Scaling attack [3,11] | 100 | 20 | 51.26 |
| GAMA-PGD [22] | 100 | 1 | 51.40 |
| MD attack [13] | 100 | 1 | 51.47 |
| PGD-50 (1000 RR) [16] | 50 | 1000 | 55.37 |
| PGD-1000 [16] | 1000 | 1 | 56.15 |

Table 8. **Evaluation against an ensemble of AutoAttack (AA) [7] and Multi-Targeted (MT) attack [10]:** Evaluating against the Multi-Targeted (MT) attack along with AutoAttack [7] leads to a marginal decrease in robust accuracy, thus showing that AutoAttack [7] is sufficient to obtain a reliable estimate of robustness.

| Method | Clean Accuracy | AA (8/255) | MT (8/255) | AA + MT (8/255) | SQ + RS (16/255) |
|---|---|---|---|---|---|
| **CIFAR-10, WRN-34-10** | | | | | |
| AWP | **85.36** | 56.17 | 56.17 | 56.15 | 30.87 |
| Ours | 85.32 | **58.04** | **58.06** | **58.03** | **35.31** |
| **CIFAR-100, WRN-34-10** | | | | | |
| AWP | 62.73 | 29.92 | 29.92 | 29.91 | 14.96 |
| Ours | **65.73** | **30.35** | **30.49** | **30.34** | **17.15** |

Further, we evaluate the model on PGD 50-step attack run with 1000 restarts. The robust accuracy saturates with increasing restarts, with the final accuracy still being higher than that achieved on AutoAttack. Lastly, we observe that the PGD-1000 attack is not very strong, confirming that the accuracy does not continually decrease as the number of steps used in the attack increases. Thus,

we observe that the proposed approach is robust against a diverse set of attack methods, thereby confirming the absence of gradient masking and verifying that the model is truly robust.

We also evaluate the WideResNet-34-10 model trained using OA-AT (proposed approach) and AWP [28] on CIFAR-10 and CIFAR-100 datasets, against an ensemble of AutoAttack [7] and Multi-Targeted attacks [10] in Table-8. We observe that using Multi-Targeted attack along with AutoAttack only leads to a drop of 0.01-0.02 % in the robust accuracy, suggesting that AutoAttack [7] is sufficient to obtain a reliable estimate of robustness.

## 9 Details on Contrast Calculation

In order to determine the contrast level for a given image, the mean absolute deviation of each pixel is first computed for the three RGB color channels independently. Following this, top 20% of pixels which correspond to the highest mean absolute deviations averaged over the three channels are selected. The variance in intensities over these selected pixels, averaged over the three channels, is used as a measure of contrast for the image. We sort images in order of increasing contrast and split the dataset into 10 bins for the evaluations in Fig.5 of the Main paper. We present the Low and High Contrast images on SVHN, CIFAR-10 and CIFAR-100 datasets respectively in Fig.10, 11, 12, 13, 14 and 15.

## 10 Sensitivity towards Hyperparameters

We check the sensitivity of the proposed method across variation in different hyperparameters on the CIFAR-10 dataset with ResNet-18 model architecture using a 110 epoch training schedule. The value of the mixup coefficient is varied from 0.6 to 1 as seen in Fig.7. On increasing the value of mixup coefficient, clean accuracy drops due to the presence of Oracle-Sensitive adversarial examples. While a lower value of mixup coefficient helps in improving clean accuracy, it makes the attack weaker, resulting in a lower robust accuracy. We visualize the effect of changing the maximum value of LPIPS coefficient in Fig.8. Using a higher LPIPS coefficient helps in boosting the clean accuracy while dropping the adversarial accuracy, while a low value close to zero drops both clean as well as robust accuracy due to the presence of oracle sensitive examples. Finally, we show the effect of changing the $\varepsilon$ (referred to as $\varepsilon_{ref}$) used in the mixup iteration. We find that using a higher value of $\varepsilon_{ref}$ in mixup iteration leads to weak attack since we project every perturbation to a much lower epsilon value while training, resulting in a higher clean accuracy and lower robust accuracy. However, a higher value of $\varepsilon_{ref}$ also leads to a more reliable estimate of the Oracle prediction, thereby leading to improved robust accuracy at intermediate values of $\varepsilon_{ref}$. Overall, we observe that OA-AT is less sensitive to hyperparameter changes.

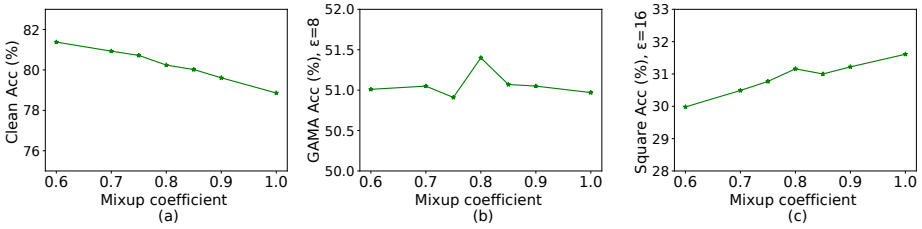


Fig. 7. **Sensitivity of the proposed approach against variation in Mixup coefficient:** (a) Clean Accuracy (%), (b) Accuracy (%) against GAMA-PGD 100-step attack [22] at $\varepsilon = 8/255$ and (c) Accuracy (%) against Square Attack [1] at $\varepsilon = 16/255$ are reported on the CIFAR-10 dataset for ResNet-18 architecture. The optimal setting chosen is mixup coefficient of 0.8.

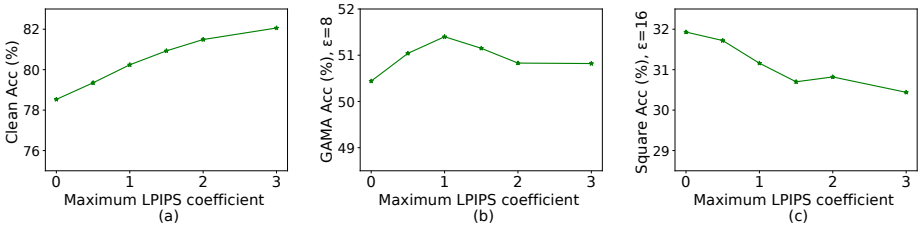


Fig. 8. **Sensitivity of the proposed approach against variation in Maximum LPIPS coefficient:** (a) Clean Accuracy (%), (b) Accuracy (%) against GAMA-PGD 100-step attack [22] at $\varepsilon = 8/255$ and (c) Accuracy (%) against Square Attack [1] at $\varepsilon = 16/255$ are reported on the CIFAR-10 dataset for ResNet-18 architecture. The optimal setting chosen is maximum LPIPS coefficient of 1.

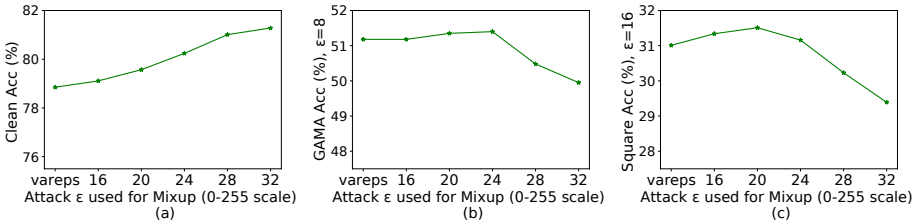


Fig. 9. **Sensitivity of the proposed approach against variation in $\varepsilon$ used in mixup iteration:** (a) Clean Accuracy (%), (b) Accuracy (%) against GAMA-PGD 100-step attack [22] at $\varepsilon = 8/255$ and (c) Accuracy (%) against Square Attack [1] at $\varepsilon = 16/255$ are reported on the CIFAR-10 dataset for ResNet-18 architecture. The optimal setting chosen is $\varepsilon = 24$ for mixup.

**Transferability of hyperparameters across datasets:** Although the proposed approach introduces two additional hyperparameters ($\alpha$ and $\lambda$), we show in Table-9 that even if we use the hyperparameters fine-tuned on CIFAR-10

dataset (WRN-34-10), they work well on SVHN (PreActResNet-18) and CIFAR-100 (WRN-34-10) datasets as well, thus showing good performance even without any fine-tuning. The gains obtained after fine-tuning specifically for the dataset are marginal.

Table 9. **Transferability of hyperparameters across datasets:** Performance (%) of the proposed method using different sets of training hyperparameters when compared to the AWP [28] baseline. The setting, *tuned* indicates that the hyperparameters have been specifically tuned for the given dataset, while *no tuning* indicates that we use the same set of hyperparameters that were found on a different dataset (CIFAR-10). The performance of the *tuned* case is only marginally better than the *no tuned* case indicating that the proposed method is not sensitive to changes in hyperparameters.

| **Method** | **Clean** | **GAMA (4/255)** | **Square (12/255)** | **GAMA (8/255)** |
|---|---|---|---|---|
| | **SVHN, PreActResNet18** | | | |
| AWP | 91.91 | 75.92 | 35.78 | 53.88 |
| Ours (tuned) | 94.61 | 78.37 | 39.56 | 55.15 |
| Ours (no tuning) | 95.17 | 78.16 | 39.12 | 54.77 |
| **Method** | **Clean** | **GAMA (8/255)** | **Square (16/255)** | **GAMA (16/255)** |
| | **CIFAR100, WRN-34-10** | | | |
| AWP | 62.73 | 29.92 | 15.85 | 11.55 |
| Ours (tuned) | 65.73 | 30.90 | 18.47 | 13.21 |
| Ours (no tuning) | 64.66 | 31.18 | 17.93 | 12.93 |

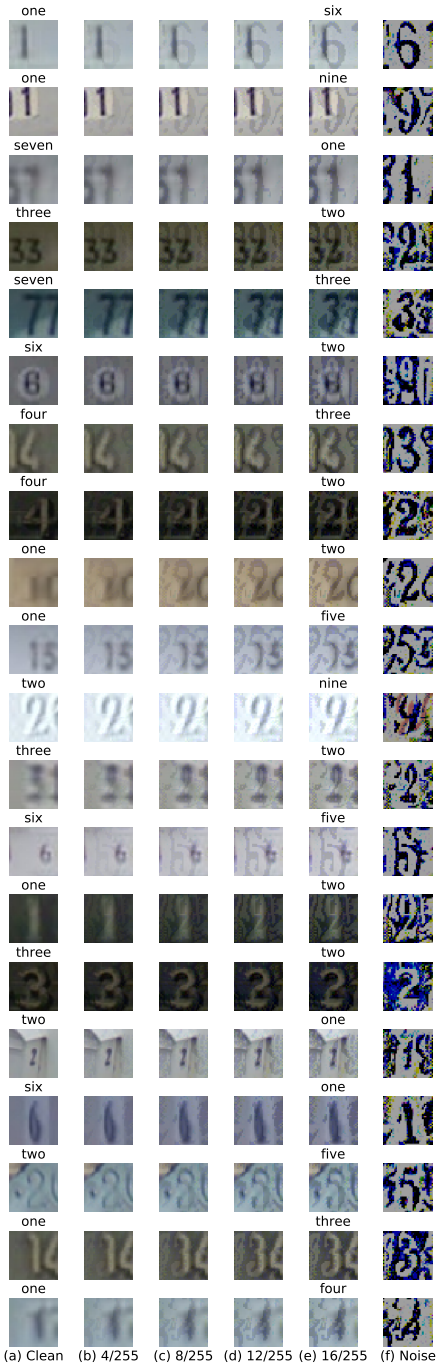


Fig. 10. **SVHN, Low-Contrast**

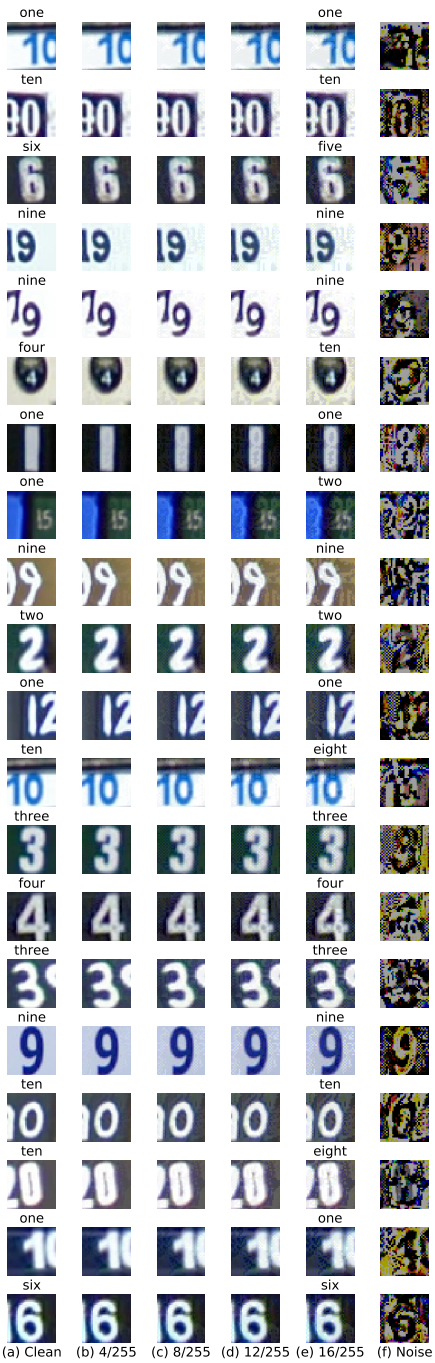


Fig. 11. **SVHN, High Contrast**

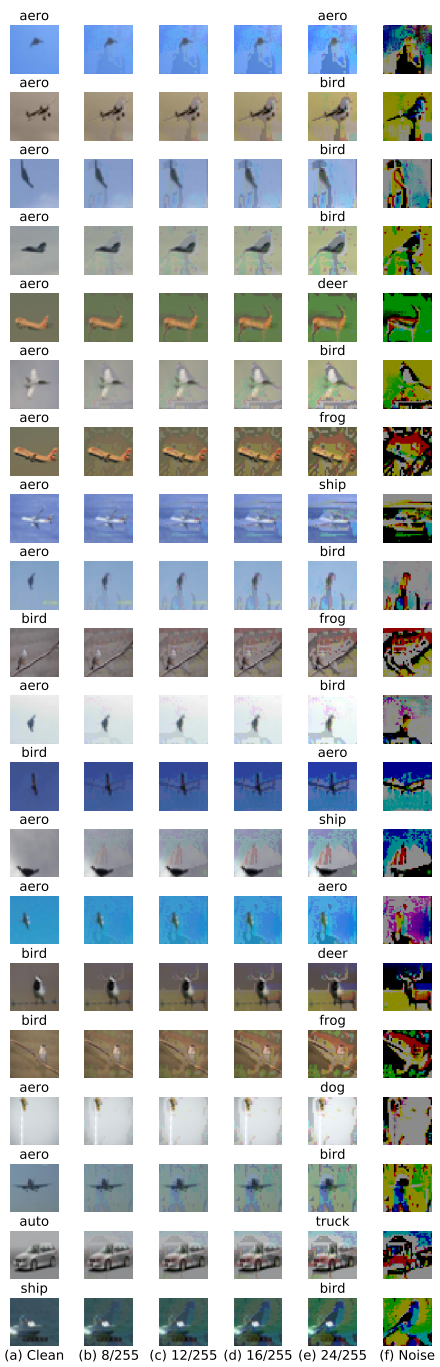


Fig. 12. **CIFAR-10 Low Contrast**

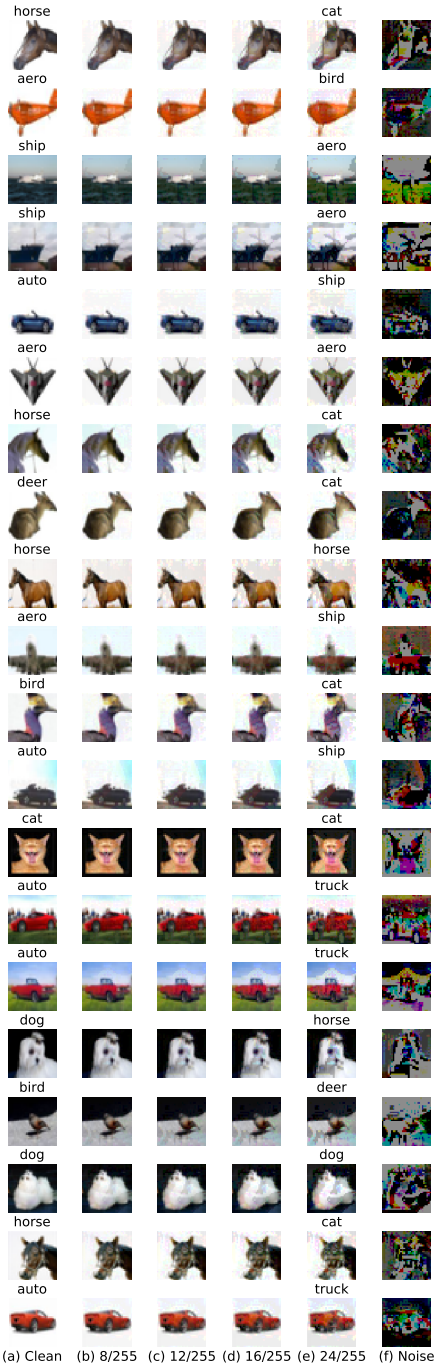


Fig. 13. **CIFAR-10 High Contrast**

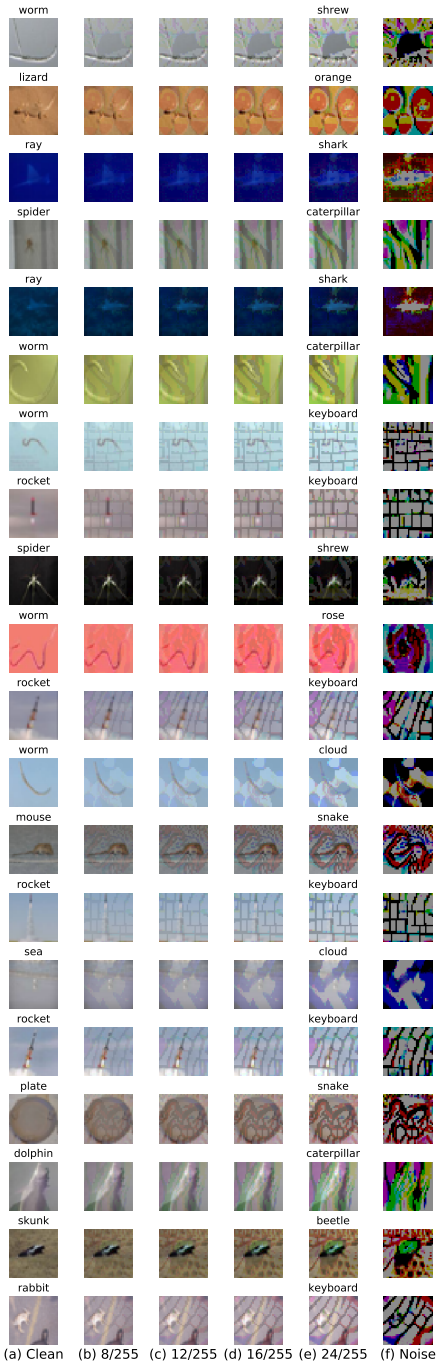


Fig. 14. **CIFAR-100 Low Contrast**

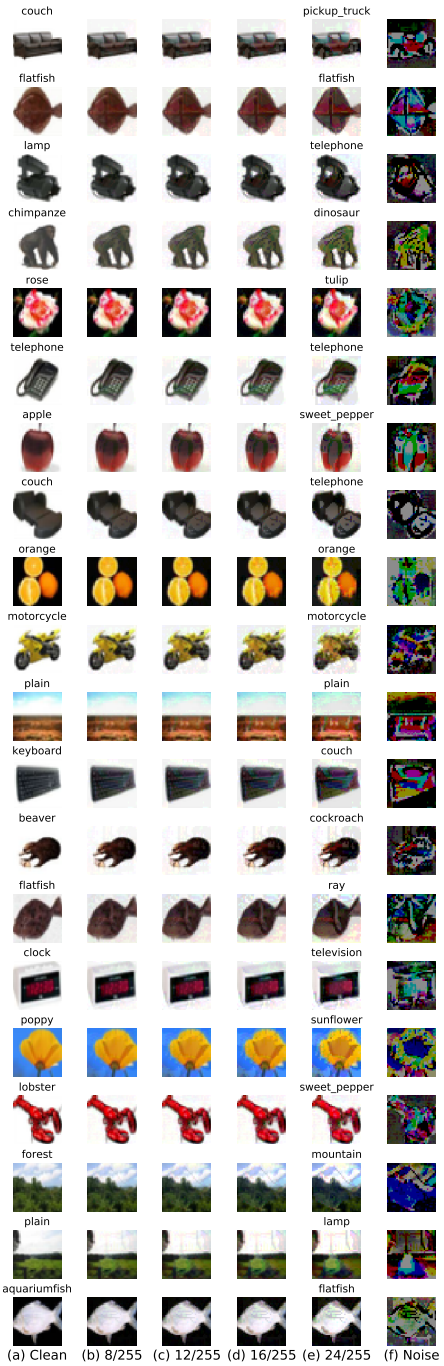


Fig. 15. **CIFAR-100 High Contrast**

## 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) 1, 2, 8, 10, 13, 14, 17
2. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International Conference on Machine Learning (ICML) (2018) 12, 13
3. Carlini, N., Wagner, D.: Defensive distillation is not robust to adversarial examples. arXiv preprint arXiv:1607.04311 (2016) 13, 15
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) 1, 2, 8
5. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM workshop on artificial intelligence and security. pp. 15–26 (2017) 8
6. Croce, F., Hein, M.: Minimally distorted adversarial examples with a fast adaptive boundary attack. In: International Conference on Machine Learning (ICML) (2020) 8, 13
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) 10, 11, 13, 15, 16
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) (June 2019) 7, 9, 11, 12, 14
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2015) 10, 13
10. 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) 9, 15, 16
11. Hitaj, D., Pagnotta, G., Masi, I., Mancini, L.V.: Evaluating the robustness of geometry-aware instance-reweighted adversarial training. arXiv preprint arXiv:2103.01914 (2021) 13, 15
12. 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) 9, 12, 13, 14
13. Jiang, L., Ma, X., Weng, Z., Bailey, J., Jiang, Y.G.: Imbalanced gradients: A new cause of overestimated adversarial robustness. arXiv preprint arXiv:2006.13726 (2020) 13, 15
14. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images (2009) 6
15. Laidlaw, C., Singla, S., Feizi, S.: Perceptual adversarial robustness: Defense against unseen threat models. International Conference on Learning Representations (ICLR) (2021) 3
16. 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) 2, 3, 8, 10, 11, 15

17. 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) 6
18. Pang, T., Yang, X., Dong, Y., Su, H., Zhu, J.: Bag of tricks for adversarial training. International Conference on Learning Representations (ICLR) (2021) 2, 3
19. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the ACM Asia Conference on Computer and Communications Security (ACM ASIACCS) (2017) 8
20. Shaeiri, A., Nobahari, R., Rohban, M.H.: Towards deep learning models resistant to large perturbations. arXiv preprint arXiv:2003.13370 (2020) 10, 11
21. Sitawarin, C., Chakraborty, S., Wagner, D.: Improving adversarial robustness through progressive hardening. arXiv preprint arXiv:2003.09347 (2020) 10, 11
22. 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) 8, 10, 13, 14, 15, 17
23. 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) (October 2021) 11, 12
24. Tashiro, Y., Song, Y., Ermon, S.: Diversity can be transferred: Output diversification for white-and black-box attacks. Advances in Neural Information Processing Systems (NeurIPS) (2020) 13, 15
25. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. In: International Conference on Learning Representations (ICLR) (2018) 10
26. 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) 5
27. Uesato, J., O'Donoghue, B., Kohli, P., van den Oord, A.: Adversarial risk and the dangers of evaluating against weak attacks. In: International Conference on Machine Learning (ICML) (2018) 8
28. Wu, D., Xia, S.T., Wang, Y.: Adversarial weight perturbation helps robust generalization. Advances in Neural Information Processing Systems (NeurIPS) (2020) 2, 4, 5, 9, 10, 11, 12, 13, 14, 16, 18
29. 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) 7, 10, 11
30. 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) 10, 11
31. 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) 3