TOWARDS ACHIEVING ADVERSARIAL ROBUSTNESS BEYOND PERCEPTUAL LIMITS

Anonymous authors
Paper under double-blind review

ABSTRACT

The vulnerability of Deep Neural Networks to Adversarial Attacks has fuelled research towards building robust models. While most Adversarial Training algorithms aim towards defending attacks constrained within low magnitude ℓ_p 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_∞ bound of $16/255$ on CIFAR-10) while outperforming existing defenses (AWP, TRADES and PGD-AT) at standard perturbation 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 (Szegedy et al., 2013). 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 (which we refer to as the 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 (Goodfellow et al., 2015; Carlini et al., 2019). For example, attacks constrained within an ℓ_∞ norm 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 (Goodfellow & Papernot), 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 (Tramèr et al., 2020), 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 (Laidlaw et al., 2021). However, as noted by Tramèr et al. (2020), finding a perceptually aligned metric is as challenging as building a network that can replicate oracle predictions. Thus, it is crucial to investigate adversarial robustness using the well-defined ℓ_p norm metric under larger perturbation bounds.

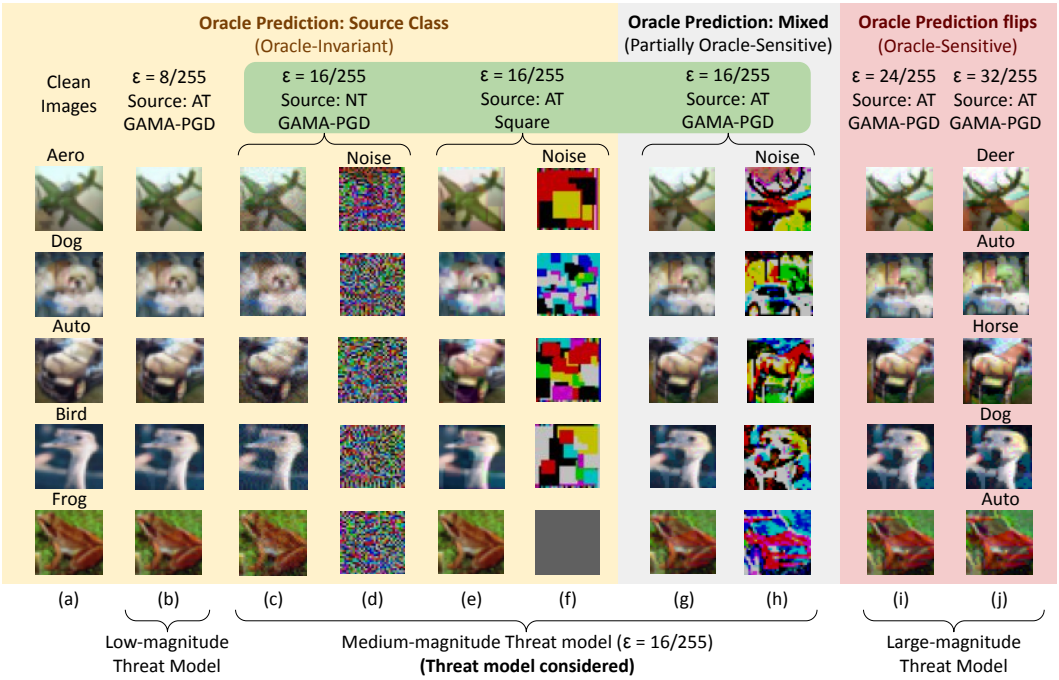


Figure 1: **Perturbations within different threat models:** Adversarially attacked images (b, c, e, g, i, j) and perturbations (d, f, h) along with the corresponding clean image (a) for various ℓ_∞ norm bounds on CIFAR-10. Attacks are generated either from an Adversarially Trained model (AT) or a Normally Trained model (NT) using the gradient-based attack GAMA-PGD (Sriramanan et al., 2020) or the Random-search based attack Square (Andriushchenko et al., 2020). The medium-magnitude threat model consists of attacks which are Oracle-Invariant and partially Oracle-Sensitive, making it a challenging setting to achieve robustness.

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 (Krizhevsky et al., 2009). 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 unperturbed image.

Our contributions have been summarized below:

- We define the ideal goals for a moderate-ε threat model (such as ℓ_∞ radius of $16/255$ for CIFAR-10 and CIFAR-100) and construct our goals as a feasible subset of the same.
- We propose methods for generating Oracle-Aligned adversaries, which can be used for adversarial training.
- 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 (Wu et al., 2020), TRADES (Zhang et al., 2019) and PGD-AT (Madry et al., 2018) at $\varepsilon = 16/255$ while also performing better at $\varepsilon = 8/255$ on CIFAR-10 and CIFAR-100. We demonstrate improved performance on SVHN (Netzer et al., 2011) as well.
- We achieve improvements over the baselines even at larger model capacities such as ResNet-34 and WideResNet-34-10.
- We empirically 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. We further show that the difference in contrast levels of images in a dataset leads to degraded robustness-accuracy trade-off.

Table 1: **CIFAR-10: Standard Adversarial Training using Large-ε perturbations** results in poor clean accuracy. Performance (%) of various existing Adversarial Defenses trained using $\varepsilon = 8/255$ or $16/255$ against attacks bound within $\varepsilon = 8/255$ and $16/255$. Defenses reported are TRADES (Zhang et al., 2019), AWP (Wu et al., 2020), PGD-AT (Madry et al., 2018) and FAT (Zhang et al., 2020).

Method	Attack ε (Training)	Clean	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

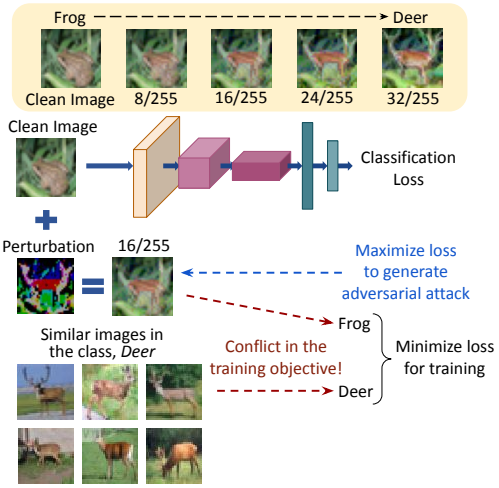


Figure 2: Issues with Standard Adversarial Training at large perturbation bounds

2 Related Works

Robustness against imperceptible attacks: Following the discovery of adversarial examples by Szegedy et al. (2013), 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. (2018) 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 (Zhang et al., 2019; 2020; Rice et al., 2020; Wu et al., 2020; Sriramanan et al., 2020; Pang et al., 2021) that improved accuracy against such imperceptible threat models further. Zhang et al. (2019) 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. (2020) 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 (Stutz et al., 2021), 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 (Izmailov et al., 2018) is also seen to improve the flatness of loss surface, resulting in a boost in adversarial robustness (Gowal et al., 2020; Chen et al., 2020). Recent works attempt to use training techniques such as early stopping (Rice et al., 2020), optimal weight decay (Pang et al., 2021), Cutmix data augmentation (Yun et al., 2019; Rebuffi et al., 2021) and label smoothing (Szegedy et al., 2016; Rebuffi et al., 2021) to achieve enhanced robust performance on base defenses such as PGD-AT (Madry et al., 2018) and TRADES (Zhang et al., 2019). We utilize some of these methods in our approach (Sec.6), and also present improved baselines by combining the strongest defense, AWP-TRADES (Wu et al., 2020) with these enhancements.

Robustness against large perturbation attacks: Shaeiri et al. (2020) 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 merely 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) (Zhang et al., 2020) 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. On similar lines, Sitawarin et al. (2020) 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).

Evaluation of Adversarial Defenses: Gradient-based white-box attacks such as PGD (Madry et al., 2018), GAMA-PGD (Sriramanan et al., 2020) and Auto-PGD with Cross-Entropy (CE) and Difference of Logits Ratio (DLR) losses (Croce & Hein, 2020a) are known to be the strongest attacks against standard Adversarial Training methods that do not obfuscate gradients. Gradient-Free attacks such as ZOO (Chen et al., 2017), SPSA (Uesato et al., 2018), Square (Andriushchenko et al., 2020) and RayS (Chen & Gu, 2020) 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 (Papernot et al., 2017). 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 additionally perform evaluations against various adaptive attacks (Sec.F) to ensure an accurate estimation of robustness for the proposed defense.

3 Preliminaries and Threat Model

In this section, we first discuss the notation used in the paper, and the nomenclature we would use for different types of adversarial attacks. We further describe the ideal goals for achieving robustness within a moderate-magnitude perturbation bound. Based on these goals, we present the objectives considered in this work and our evaluation criteria.

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 (Goodfellow & Papernot). 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 $\mathcal{S}(x) = \{\hat{x} : ||\hat{x} - x||_p < \varepsilon\}$. The set of all ℓ_p norm bound adversarial examples is defined as $\mathcal{A}(x) = \{\widetilde{x} : f_\theta(\widetilde{x}) \neq y, \widetilde{x} \in \mathcal{S}(x)\}$. 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

Tramèr et al. (2020) discuss the existence of two types of adversarial examples: Sensitivity-based examples, where the model prediction changes, but 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 adversarial examples, while models which are overly robust to large perturbation bounds could be susceptible to invariance-based examples. Since these definitions are dependent on the model being considered, we define a different nomenclature which only depends on the input image and the threat model considered, as below:

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

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

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

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

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 defined by $\varepsilon = 8/255$ 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 Out-of-Distribution (OOD) images

We incorporate these goals in the training objective of our proposed defense, which is discussed in Sec.4. Given the practical difficulty in assigning Oracle labels, we consider the following criteria for our defense evaluations:

- Robustness-Accuracy trade-off, measured using accuracy on clean samples and robustness against valid attacks within the threat model (discussed below)
- Robustness against all attacks within an imperceptible radius ($\varepsilon = 8/255$ for CIFAR-10), measured using strong white-box attacks (Croce & Hein, 2020b; Sriramanan et al., 2020)
- Robustness to Oracle-Invariant samples within a larger radius ($\varepsilon = 16/255$ for CIFAR-10), measured using gradient-free attacks (Andriushchenko et al., 2020; Chen & Gu, 2020)

In Sec.5 we show the relation between the contrast of an image and the existence of Oracle-Sensitive attacks within a given radius, and introduce additional evaluation metrics based on this.

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 (Tsipras et al., 2019). Adversarial examples generated using a robust model tend to start looking 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.

Generation of Oracle-Invariant examples: While the strongest Oracle-Invariant examples are generated using the gradient-free attacks Square (Andriushchenko et al., 2020) and Ray-S (Chen &

Figure 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.

Gu, 2020), 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 (Zhang et al., 2018; Laidlaw et al., 2021). As shown in Fig.8, while the standard AlexNet model used in prior work (Laidlaw et al., 2021) fails to distinguish between Oracle-Invariant and Oracle-Sensitive samples, an adversarially trained model is able to distinguish between the two effectively. 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. As shown in Fig.10, $\lambda = 1$ generates attacks which are Oracle-Invariant and strong on the CIFAR-10 dataset. The value of λ is further fine-tuned during training to achieve the optimal robustness-accuracy trade-off.

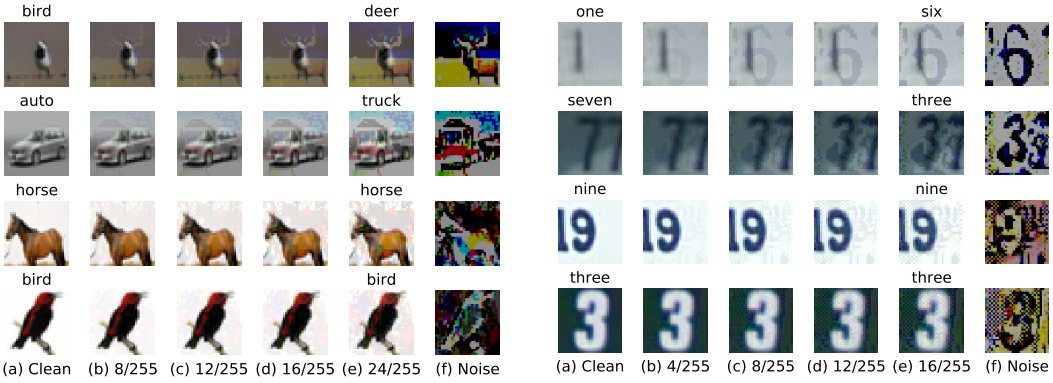
Oracle-Aligned Adversarial Training (OA-AT): The training algorithm for the proposed defense, Oracle-Aligned Adversarial Training (OA-AT) is presented in Algorithm-1 and illustrated in Fig.3. We explain the proposed algorithm by considering an example of the moderate-magnitude threat model of $\varepsilon = 16/255$ on the CIFAR-10 and CIFAR-100 datasets below.

We use the TRADES-AWP formulation (Zhang et al., 2019; Wu et al., 2020) as the base implementation, with Cross-Entropy loss instead of KL-divergence loss for attack generation, as it results in stronger attacks Gowal et al. (2020). We maximize loss on $x_i + 2 \cdot \widetilde{\delta}_i$ (where $\widetilde{\delta}_i$ is the attack) in the additional weight perturbation step, as it results in improved robust generalization. We start with an initial ε value of $4/255$ upto one-fourth the training epochs, and ramp up this value linearly to a value of $\varepsilon_{max} = 16/255$ at the last epoch alongside a cosine learning rate schedule. We use 5 attack steps when $\varepsilon = 4/255$ and 10 attack steps later.

We perform standard adversarial training upto $\varepsilon = 12/255$ as the attacks in this range are imperceptible. Beyond this, we start incorporating separate training losses for Oracle-Invariant and Oracle-Sensitive samples in alternate training iterations as shown in Fig.3. Oracle-Sensitive samples are generated by maximizing Cross-Entropy 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 at a larger bound, ε_{ref} as shown below:

$$\mathcal{L}_{adv} = KL\big(f_\theta(x_i + \widetilde{\delta}_i) \;||\; \alpha\; f_\theta(x_i) + (1-\alpha)\; f_\theta(x_i + \widehat{\delta}_i)\big) \tag{3}$$

Here $\widetilde{\delta}_i$ is the perturbation at the varying epsilon value $\widetilde{\varepsilon}$, and $\widehat{\delta}_i$ is the perturbation at ε_{ref}. We set the value of ε_{ref} to be greater than or equal to ε_{max} This results in better robustness-accuracy trade-off as shown in Table-4. In the alternate iteration, we use the LPIPS metric to efficiently generate strong Oracle-Invariant attacks during training. We perform exponential weight-averaging of the network being trained and use this for computing the LPIPS metric for improved and stable results (Table-4). We increase α and λ over training, as the nature of attacks changes with varying



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

Figure 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.

Table 2: **Adversarial Training on Contrast-Enhanced (CE) datasets:** Performance (%) of the AWP-TRADES defense (Wu et al., 2020) by performing the training and evaluation on contrast-enhanced (CE) datasets when compared to standard datasets. Evaluation is done against AutoAttack (Croce & Hein, 2020b) at different perturbation bounds. The contrast of a dataset plays a significant role in the Robustness-Accuracy trade-off achieved.

(a) **Training on $\varepsilon = 8/255$ perturbations**

Dataset	Clean	AA 4/255	AA 8/255	AA 12/255	AA 16/255
SVHN	91.91	75.72	30.31	30.31	14.37
SVHN-CE	**94.61**	**87.89**	**80.23**	**69.25**	**56.41**
CIFAR-10	80.47	66.82	49.87	33.17	19.23
CIFAR-10-CE	**82.18**	**70.70**	**57.58**	**43.04**	**28.82**
CIFAR-100	58.81	39.01	25.30	14.71	8.29
CIFAR-100-CE	**59.04**	**41.40**	**26.95**	**15.66**	**8.97**

(b) **Training on larger magnitude perturbations**

Dataset	Training epsilon	Clean	AA 4/255	AA 8/255	AA 12/255	AA 16/255
CIFAR-10	8/255	80.47	66.82	49.87	33.17	19.23
CIFAR-10-CE	8/255	**82.18**	**70.70**	**57.58**	**43.04**	**28.82**
CIFAR-10	10/255	80.32	65.70	48.89	32.61	18.81
CIFAR-10-CE	10/255	**82.05**	**71.13**	**57.81**	**43.02**	**28.82**
CIFAR-10	16/255	71.63	56.31	40.55	26.31	15.42
CIFAR-10-CE	16/255	**78.47**	**67.76**	**55.77**	**42.89**	**30.70**

$\widetilde{\varepsilon}$. 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 Role of Image Contrast in Robust Training and 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)) flip the Oracle prediction. As noted by Balaji et al. (2019), the perturbation radius at which the Oracle prediction flips 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 the top 20 Low-Contrast and High-Contrast images in the SVHN, CIFAR-10 and CIFAR-100 datasets in Fig.15,16,17,18,19,20 of the Appendix, and show a few images in Fig.4 as well. We observe that High-contrast (HC) images are Oracle-Invariant even at large perturbation bounds, whereas Low-Contrast (LC) images are Oracle-Sensitive at lower perturbation bounds as well. Based on this, we utilize High-Contrast images for evaluation against strong White-Box attacks at large epsilon bounds in Sec.6. As shown in Fig.2, the presence of Oracle-Sensitive images in the considered perturbation bound causes a drop in clean accuracy due to a conflict in training objective,

Table 3: **Comparison with existing methods:** Performance (%) of the proposed defense OA-AT when compared to baselines against the attacks, GAMA-PGD100 (Sriramanan et al., 2020), AutoAttack (AA) (Croce & Hein, 2020b) and an ensemble of Square (Andriushchenko et al., 2020) and Ray-S (Chen & Gu, 2020) attacks (SQ+RS), with different ε bounds. Sorted by AutoAttack (AA) accuracy at $\varepsilon = 8/255$ for CIFAR-10 and CIFAR-100, and $\varepsilon = 4/255$ for SVHN.

(a) **CIFAR-10**

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), 110 epochs						
AWP	85.19	55.87	55.69	31.27	24.04	23.46
AWP+	85.10	56.07	55.87	31.36	23.79	23.27
OA-AT (Ours)	**85.67**	**56.45**	**55.93**	**33.89**	**25.21**	**24.05**

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

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**
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**

hinting at the fact that contrast levels of images in the dataset can impact the accuracy-robustness trade-off significantly. We validate this hypothesis by performing standard adversarial training using the AWP-TRADES algorithm (Wu et al., 2020) and robust evaluation using AutoAttack (Croce & Hein, 2020a) on a contrast-enhanced dataset (contrast enhancement is done for both train and test set) that is generated using histogram equalization. We present results in Table-2a. We observe significant gains of around 40% or higher against AutoAttack at large ε values on SVHN. We observe noteworthy gains on the CIFAR-10 and CIFAR-100 datasets as well. Since the SVHN dataset has the highest imbalance in terms of contrast levels of images, this dataset benefits the most with histogram equalization. We note that this is not a practical option for improving the robustness of models as images in the test set would be of varying contrast levels and their contrast levels before attack generation cannot be controlled. Moreover, incorporating pre-processing methods as part of the defense are known to be susceptible to adaptive attacks (Athalye et al., 2018; Carlini et al., 2019; Tramer et al., 2020) that consider the defense strategy to generate stronger attacks. This experiment merely highlights the role of contrast in robustness-accuracy trade-off in adversarial training. We further show in Table-2b that the contrast-enhanced dataset can be used for adversarial training at relatively larger perturbation bounds as well, without a significant drop in clean accuracy.

6 Experiments and Results

We compare performance of the proposed approach with the existing defenses discussed in Sec.2 on the CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) using the ResNet-18 architecture for a training budget of 110 epochs. On CIFAR-10, for each baseline, we find the best set of hyperparameters to achieve clean accuracy of around 80% to ensure a fair comparison across all methods. We also perform baseline training across various ε values and report the best baselines in Table-3a. We observe that baseline defenses do not perform well when trained using large ε bounds such as $16/255$ as shown in Table-1 (Detailed results in Table-5). We compare the proposed approach against the strongest baseline AWP-TRADES (Wu et al., 2020) on CIFAR-100 in Table-3a and show a more detailed comparison against more baselines in Table-6.

Based on the baseline evaluations on CIFAR-10 and CIFAR-100 datasets, we find that the strongest baseline is AWP-TRADES, which we refer to as AWP in the tables. We therefore compare the proposed approach against this baseline for the CIFAR-10 and CIFAR-100 evaluations on other

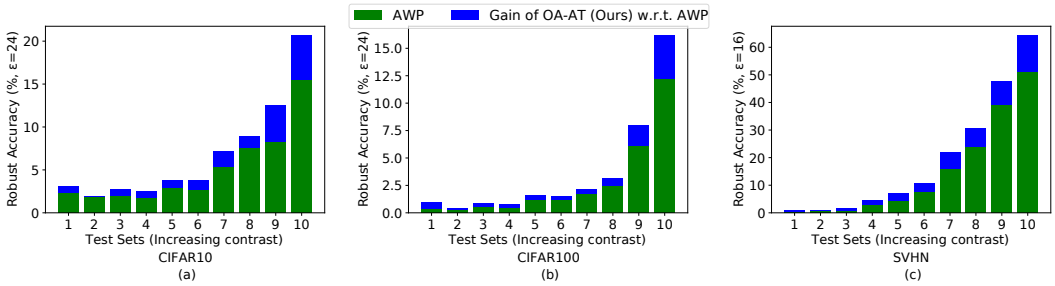


Figure 5: **Evaluation across test subsets of increasing contrast levels:** Performance comparison of the proposed defense OA-AT against AWP-Trades (Wu et al., 2020) across test subsets of increasing contrast levels. The proposed defense achieves higher gains as the contrast of the images in the test subsets increases, verifying that the proposed approach is more robust to the Oracle-Invariant white-box attacks on High-Contrast images.

model architectures (ResNet-34, WideResNet-34-10 and PreActResNet-18), and for evaluations on the SVHN dataset. We observe additional gains with the use of techniques such as AutoAugment (Cubuk et al., 2018; Stutz et al., 2021; Rebuffi et al., 2021) and Model Weight Averaging (WA) (Izmailov et al., 2018; Gowal et al., 2020; Chen et al., 2020), 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-3 (Detailed results in Table-8). As observed by Rebuffi et al. (2021), 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.

On the CIFAR-10 and CIFAR-100 datasets, we report adversarial robustness against the strongest known attacks, AutoAttack (AA) (Croce & Hein, 2020b) and GAMA PGD-100 (GAMA) (Sriramanan et al., 2020) 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 (Andriushchenko et al., 2020) and Ray-S (Chen & Gu, 2020) attacks, as they are the strongest known Oracle-Invariant attacks. On the SVHN dataset, we find that the perturbation bound for imperceptible attacks is $\varepsilon = 4/255$, and we consider robustness within $\varepsilon = 12/255$. The proposed defense achieves consistent gains across all metrics considered in Sec.3.3. 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 of the existing methods (Table-5). As show in Fig.5, the proposed defense achieves higher gains on the high contrast subsets of all datasets, verifying that the proposed approach has higher gains in robustness against Oracle-Invariant attacks, and not against Oracle-Sensitive attacks. We further evaluate the proposed defense against diverse attacks (Table-7) and sanity checks (Sec.F) to ensure the absence of gradient masking.

7 Conclusions

We explore the idea of robustness beyond perceptual limits in an ℓ_p norm based threat model. We first discuss the ideal goals of an adversarial defense at larger perturbation bounds, and further propose a novel defense, Oracle-Aligned Adversarial Training (OA-AT) that aims to align 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 significant gains in 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 to highlight the role of contrast-level 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.

8 REPRODUCIBILITY STATEMENT

We share the code for reproducing the results of the proposed method OA-AT along with the Supplementary submission.

REFERENCES

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *The European Conference on Computer Vision (ECCV)*, 2020.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018.

Yogesh Balaji, Tom Goldstein, and Judy Hoffman. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *arXiv preprint arXiv:1910.08051*, 2019.

Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Jinghui Chen and Quanquan Gu. 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.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 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.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations (ICLR)*, 2020.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning (ICML)*, 2020a.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020b.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Ian Goodfellow and Nicolas Papernot. Is attacking machine learning easier than defending it? Blog post on Feb 15, 2017.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

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

Dorjan Hitaj, Giulio Pagnotta, Iacopo Masi, and Luigi V Mancini. Evaluating the robustness of geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2103.01914*, 2021.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Linxi Jiang, Xingjun Ma, Zejia Weng, James Bailey, and Yu-Gang Jiang. Imbalanced gradients: A new cause of overestimated adversarial robustness. *arXiv preprint arXiv:2006.13726*, 2020.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *International Conference on Learning Representations (ICLR)*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Tsipras Dimitris, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *International Conference on Learning Representations (ICLR)*, 2021.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the ACM Asia Conference on Computer and Communications Security (ACM ASIACCS)*, 2017.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*, 2020.

Amirreza Shaeiri, Rozhin Nobahari, and Mohammad Hossein Rohban. Towards deep learning models resistant to large perturbations. *arXiv preprint arXiv:2003.13370*, 2020.

Chawin Sitawarin, Supriyo Chakraborty, and David Wagner. Improving adversarial robustness through progressive hardening. *arXiv preprint arXiv:2003.09347*, 2020.

Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R Venkatesh Babu. Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

David Stutz, Matthias Hein, and Bernt Schiele. Relating adversarially robust generalization to flat minima. *arXiv preprint arXiv:2104.04448*, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Yusuke Tashiro, Yang Song, and Stefano Ermon. Diversity can be transferred: Output diversification for white-and black-box attacks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.

Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *International Conference on Machine Learning (ICML)*, 2020.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning*, pp. 11278–11287. PMLR, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

A ORACLE-INVARIANT ATTACKS

Square Attack: The strongest Oracle-Invariant examples are generated using the Square attack (Andriushchenko et al., 2020). 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 6.

RayS Attack: Another technique that is observed to generate strong Oracle-Invariant examples is the black-box RayS attack (Chen & Gu, 2020). 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 7. **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 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 λ_X in each case. The value of λ_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.

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:

$$\mathcal{L}_{CE}(x, y) - \lambda_{Disc} \cdot \mathcal{L}_{BCE}(\hat{x}, \text{OI}) \tag{4}$$

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

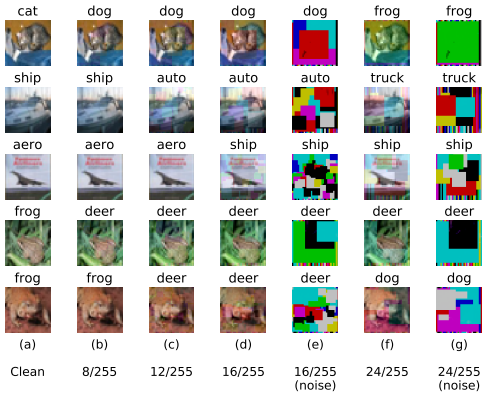


Figure 6: **Square attack:** Adversarially attacked images (b, c, d, f) and the corresponding perturbations (e, g) for various ℓ_∞ bounds generated using the gradient-free random search based attack Square (Andriushchenko et al., 2020). The clean image is shown in (a). Attacks are generated from a model trained using the proposed Oracle-Aligned Adversarial Training (OA-AT) algorithm on CIFAR-10. Prediction of the same model is printed above each image.

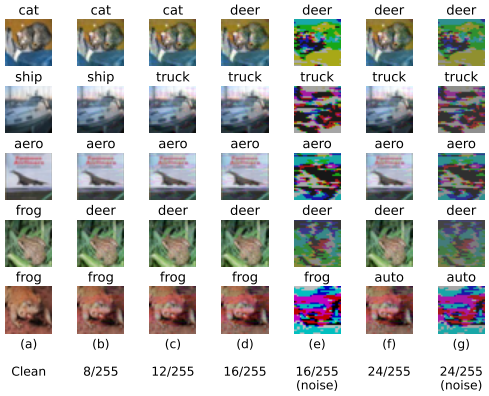


Figure 7: **RayS attack:** Adversarially attacked images (b, c, d, f) and the corresponding perturbations (e, g) for various ℓ_∞ bounds generated using the gradient-free binary search based attack RayS (Chen & Gu, 2020). The clean image is shown in (a). Attacks are generated from a model trained using the proposed Oracle-Aligned Adversarial Training (OA-AT) algorithm on CIFAR-10. Prediction of the same model is printed above each image.

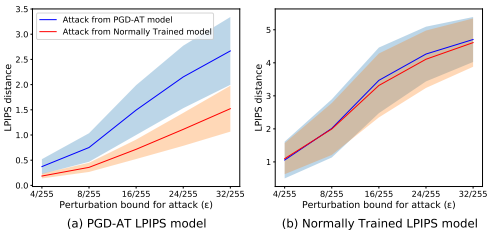


Figure 8: LPIPS distance between clean and adversarially perturbed images. Attacks generated from PGD-AT (Madry et al., 2018; Pang et al., 2021) model (Oracle-Sensitive) and Normally Trained model (Oracle-Invariant) are considered. (a) PGD-AT ResNet-18 model is used for computation of LPIPS distance (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.

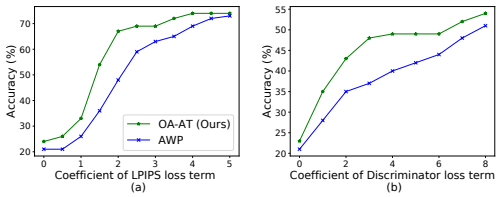


Figure 9: Comparison of the proposed model with the best baseline, AWP (Wu et al., 2020) trained on CIFAR-10 with ResNet-18 architecture, against attacks of varying strength and Oracle sensitivity constrained within perturbation bound of $\varepsilon = 16/255$. (a) LPIPS based regularizer, and (b) Discriminator 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 (OA-AT) achieves improved accuracy compared to AWP.

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.4 therefore attempts to break the most salient property of Oracle-Sensitive attacks, which is the spatial correlation between an image and its perturbation.

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 (Zhang et al., 2018; Laidlaw et al., 2021). As shown in Fig.8, while the standard AlexNet model that is used in prior work (Laidlaw et al., 2021) fails to distinguish between

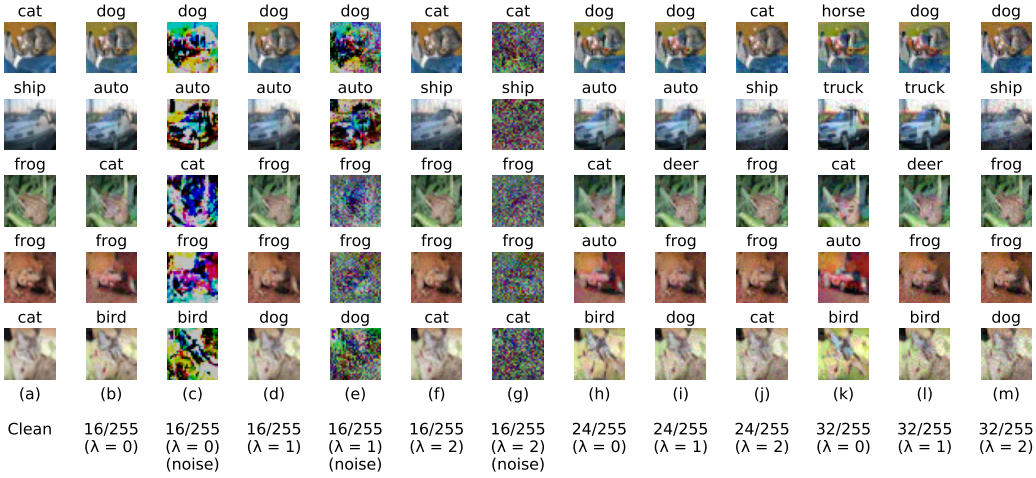


Figure 10: Oracle-Invariant adversarial examples generated using the LPIPS based PGD attack in Eq.5 across various perturbation bounds. White-box attacks and predictions on the model trained using the proposed OA-AT defense on the CIFAR-10 dataset with ResNet-18 architecture are shown: (a) Original Unperturbed 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 λ_{LPIPS} to 1 and 2, (c, e, g) Perturbations corresponding to (b), (d) and (f) respectively.

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 (Madry et al., 2018; Pang et al., 2021) model (Fig.1(c-e)) as Oracle-Sensitive attacks, and attacks generated from a Normally Trained model (Fig.1(h)) 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:

$$\mathcal{L}_{CE}(x, y) - \lambda_{\text{LPIPS}} \cdot \text{LPIPS}(x, \hat{x}) \tag{5}$$

We choose λ_{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.10, setting λ_{LPIPS} to 1 changes adversarial examples from Oracle-Sensitive to Oracle-Invariant, as they look similar to the corresponding original images shown in Fig.10(a). This can be observed more distinctly at perturbation bounds of $24/255$ and $32/255$. The perturbations in Fig.10(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.10(f, j, m) where the model prediction is same as the true label. We therefore set the value of λ_{LPIPS} to 1 to obtain strong Oracle-Invariant attacks.

As shown in Table-4, while we obtain the best results using the LPIPS based PGD attack for training (E1), the use of discriminator based PGD attack (E6) 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.

Evaluation of the proposed defense against Oracle-Invariant Attacks: We compare the performance of the proposed defense OA-AT with the strongest baseline AWP (Wu et al., 2020) against the two proposed Oracle-Invariant attacks, LPIPS based attack and Discriminator based attack in Fig.9 (a) and (b) respectively. We vary the coefficient of the regularizers used in the generation of attacks, λ_{Disc} (Eq.4) and λ_{LPIPS} (Eq.5) 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 when compared to the AWP (Wu et al., 2020) baseline.

Algorithm 1 Oracle-Aligned Adversarial Training

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
for epoch $= 1$ **to** T **do**
  $\widetilde{\varepsilon} = \max\{\varepsilon_{max}/4, \varepsilon_{max} \cdot \text{epoch}/T\}$
  for $i = 1$ **to** M **do**
    $\delta_i \sim U(-\min(\widetilde{\varepsilon}, \varepsilon_{max}/4), \min(\widetilde{\varepsilon}, \varepsilon_{max}/4))$
    if $\widetilde{\varepsilon} < 3/4 \cdot \varepsilon_{max}$ **then**
      $\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i)$
      $\widetilde{\delta}_i = A(x_i, y_i, \ell, \widetilde{\varepsilon})$
      $L_{adv} = D_{\text{KL}}\big(f_\theta(x_i + \widetilde{\delta}_i) || f_\theta(x_i)\big)$
    else if $i \,\%\, 2 = 0$ **then**
      $\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i)$
      $\widehat{\delta}_i = A(x_i, y_i, \ell, 1.5 \cdot \varepsilon_{max})$
      $\widetilde{\delta}_i = \Pi_\infty(\widehat{\delta}_i, \widetilde{\varepsilon})$
      $L_{adv} = D_{\text{KL}}\left(f_\theta(x_i + \widetilde{\delta}_i) \quad || \quad \alpha \cdot f_\theta(x_i) + (1-\alpha) \cdot f_\theta(x_i + \widehat{\delta}_i)\right)$
    else
      $\delta_i \sim U(-\widetilde{\varepsilon}, \widetilde{\varepsilon})$
      $\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i) - \text{LPIPS}(x_i, x_i + \delta_i)$
      $\widetilde{\delta}_i = A(x_i, y_i, \ell, \widetilde{\varepsilon})$
      $L_{adv} = D_{\text{KL}}\big(f_\theta(x_i + \widetilde{\delta}_i) || f_\theta(x_i)\big)$
    end if
    $L = \ell_{CE}(f_\theta(x_i), y_i) + L_{adv}$
    $\theta = \theta - \eta \cdot \nabla_\theta L$
  end for
end for

B Details on the Datasets used

We evaluate the proposed approach on the CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009) and SVHN (Netzer et al., 2011) 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 ℓ_∞ threat model of radius $8/255$ to be representative of imperceptible perturbations, that is, the Oracle label does not change within this set. While for SVHN we consider this bound to be $4/255$. Further, we consider the ℓ_∞ threat model of radius $16/255$ to investigate robustness within moderate magnitude perturbation bounds for CIFAR-10 and CIFAR-100 while this bound is $12/255$ for SVHN dataset.

C Details on Training

The algorithm for the proposed method as explained in Sec.4 is presented in Algorithm-1. We use a varying ε schedule and start training on perturbations of magnitude $\varepsilon = 4/255$. This results in marginally better performance when compared to ramping up the value of ε from 0 (E8 of Table-4). For CIFAR-10 training on ResNet-18, we set the weight of the adversarial loss L_{adv} in L21 of Alg.1 (β parameter of TRADES (Zhang et al., 2019)) 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 ε is linearly

Table 4: **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 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: Using Discriminator instead of LPIPS (OI Attack)	80.56	50.75	22.13	31.17	58.84	26.35	10.64	14.82
E7: Without 2*eps perturbations for AWP	79.96	50.50	22.61	30.60	60.18	26.27	10.15	14.20
E8: Increasing epsilon from the beginning	80.34	50.77	22.57	30.80	**60.51**	26.34	10.37	14.61
E9: Maximizing KL div in the AWP step	81.19	49.77	21.17	29.39	59.48	25.03	7.93	13.34
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: With fixed ε=16/255	71.64	47.59	25.91	31.75	50.99	23.19	9.99	13.48

increased from $12/255$ to $16/255$. In this moderate perturbation regime, we also linearly increase the coefficient of the LPIPS distance (Alg.1, L17) from 0 to 1, and linearly decrease the α parameter used in the convex combination of softmax prediction (Alg.1, L14) 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.

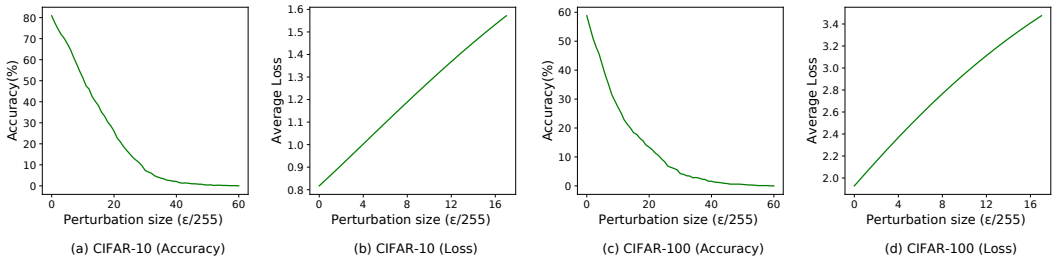
For all our experiments, we use the cosine learning rate schedule with 0.2 as the maximum learning rate. We use SGD optimizer with momentum of 0.9, and train for 110 epochs, except for training PreActResNet18 on CIFAR-100 where we use 200 epochs. We compute the LPIPS distance using an exponential weight averaged model with $\tau = 0.995$. We note from Table-4 that the use of weight-averaged model 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 (Cubuk et al., 2018) 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 robust 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-4), 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-4) to 1 (E12 in Table-4). We use AutoAugment with $p = 1$ for SVHN as we find it helps in more stable training of our method. Further we find that using Label Smoothing with CIFAR-100 helps in improving the clean accuracy as shown in Table-3.

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.

D 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-4. 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



(a) CIFAR-10 (Accuracy) (b) CIFAR-10 (Loss) (c) CIFAR-100 (Accuracy) (d) CIFAR-100 (Loss)

Figure 11: 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 (ε) 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 (Athalye et al., 2018) (b, d) Plots showing the variation of Cross-Entropy Loss on FGSM attack (Goodfellow et al., 2015) against variation in the attack perturbation bound (ε) on CIFAR-10 and CIFAR-100 datasets. As the perturbation bound increases, loss increases linearly, indicating the absence of gradient masking (Athalye et al., 2018)

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 ($\varepsilon = 16/255$), it leads to poor clean accuracy as shown in the first partition of Table-5.

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 E7. 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 E9 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. (2020). We also investigate the effect of AutoAugment (Cubuk et al., 2018), Weight Averaging (Izmailov et al., 2018) and Label Smoothing + Warmup on AWP (Wu et al., 2020) baseline in Table- 8.

E Detailed Results

In Tables-5 and 6, we present results of different defense methods such as AWP-TRADES (Wu et al., 2020), TRADES (Zhang et al., 2019), PGD-AT (Madry et al., 2018), ExAT (Shaeiri et al., 2020), ATES (Sitawarin et al., 2020) and FAT (Zhang et al., 2020), evaluated across a wide range of adversarial attacks. We present evaluations on the Black-Box FGSM attack (Goodfellow et al., 2015) and a suite of White-Box attacks, on ℓ_∞ 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 (Tramèr et al., 2018), the GAMA PGD-100 attack (Sriramanan et al., 2020) and AutoAttack (Croce & Hein, 2020b), with the latter two being amongst the strongest of attacks known to date. Lastly, we also present evaluations on the Square attack (Andriushchenko et al., 2020) for $\varepsilon = 12/255$ and $16/255$ in order to evaluate performance on Oracle-Invariant samples at large perturbation bounds.

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-5. In the first partition of the table, we present baselines trained using attacks constrained within an ℓ_∞ 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

Table 5: **CIFAR-10**: Performance (%) of the proposed defense OA-AT against attacks with different ε bounds, when compared to the following baselines: AWP (Wu et al., 2020), ExAT (Shaeiri et al., 2020), TRADES (Zhang et al., 2019), ATES (Sitawarin et al., 2020), PGD-AT (Madry et al., 2018) and FAT (Zhang et al., 2020). AWP (Wu et al., 2020) 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 (Croce & Hein, 2020b) (AA $8/255$). The proposed defense achieves significant gains in accuracy across all attacks.

Method	Attack ε (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

Table 6: **CIFAR-100**: Performance (%) of the proposed defense OA-AT against attacks with different ε bounds, when compared to the following baselines: AWP (Wu et al., 2020), ExAT (Shaeiri et al., 2020), TRADES (Zhang et al., 2019), ATES (Sitawarin et al., 2020), PGD-AT (Madry et al., 2018) and FAT (Zhang et al., 2020). AWP (Wu et al., 2020) is the strongest baseline. The baselines are sorted by AutoAttack accuracy (Croce & Hein, 2020b) (AA $8/255$). The proposed defense achieves significant gains in accuracy against the strongest attacks across all ε bounds. Since the proposed defense uses AutoAugment (Cubuk et al., 2018) as the augmentation strategy, we present results on the strongest baseline AWP (Wu et al., 2020) with AutoAugment as well.

Method	Attack ε (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)
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
OA-AT (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

attack strengths tuned 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-5. We observe that the proposed method OA-AT consistently outperforms other approaches on all three metrics described in Sec.3.3, 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 ℓ_∞ constraint set of $8/255$ when compared to existing approaches, despite being trained on larger perturbations sets.

CIFAR-100: In Table-6, 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 (Cubuk et al., 2018; Stutz et al., 2021). To enable fair comparison, we incorporate AutoAugment for the strongest baseline, AWP (Wu et al., 2020) 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.

Table 7: **Evaluation against various attacks with a perturbation bound of $\varepsilon = 8/255$ on CIFAR-10**: Performance (%) of the proposed defense OA-AT 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† Croce & Hein (2020b)	100	20	**50.88**
GAMA-MT Sriramanan et al. (2020)	100	5	50.90
ODS (98 +2 steps) Tashiro et al. (2020)	100	100	50.94
MDMT attack Jiang et al. (2020)	100	10	51.19
Logit-Scaling attack Carlini & Wagner (2016); Hitaj et al. (2021)	100	20	51.26
GAMA-PGD Sriramanan et al. (2020)	100	1	51.40
MD attack Jiang et al. (2020)	100	1	51.47
PGD-50 (1000 RR) Madry et al. (2018)	50	1000	55.37
PGD-1000 Madry et al. (2018)	1000	1	56.15

F Gradient Masking Checks

As discussed by Athalye et al. (2018), we present various checks to ensure the absence of Gradient Masking in the proposed defense. In Fig.11(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.11(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, as compared to strong gradient based attacks such as GAMA or AutoAttack in Tables-5,6. 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.

In Table-7, we perform exhaustive evaluations using various attack techniques to further verify the absence of gradient masking. In addition to AutoAttack (Croce & Hein, 2020b) which in itself consists of an ensemble of four attacks- AutoPGD with Cross-Entropy and Difference-of-Logits loss, the FAB attack (Croce & Hein, 2020a) and Square Attack (Andriushchenko et al., 2020), we present evaluations against strong multi-targeted attacks such as GAMA-MT (Sriramanan et al., 2020) and the MDMT attack (Jiang et al., 2020) 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 (Tashiro et al., 2020) 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 (Carlini & Wagner, 2016; Hitaj et al., 2021) helps yield robust evaluations that are less dependent on the exact scale of output logits predicted by the network, and is seen to be 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.

Furthermore, 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.

G 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

Table 8: **Effect of Ancillary Methods applied to AWP baseline:** Performance (%) of models trained by applying AutoAugment (Cubuk et al., 2018), Label Smoothing + Warmup and Weight Averaging (Izmailov et al., 2018) to the AWP baseline (Wu et al., 2020), against GAMA-PGD100 (Sriramanan et al., 2020) and Square (Andriushchenko et al., 2020) attacks. Results on the CIFAR-10, CIFAR-100 and SVHN datasets are reported using different ε bounds.

AutoAugment 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), 110 epochs						
0	×	×	85.19	55.87	32.68	24.04
0	×	✓	85.10	56.07	32.50	23.79
0.5	×	×	85.42	55.40	32.25	22.48
0.5	×	✓	84.85	55.00	32.26	22.50
1	×	×	84.60	51.62	29.73	19.77
1	×	✓	83.81	52.22	30.06	20.03
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	11.55	15.85
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

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.

H Sensitivity of OA-AT 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.12. On reducing the value of mixup coefficient, clean accuracy drops sharply due to the presence of Oracle-Sensitive adversarial

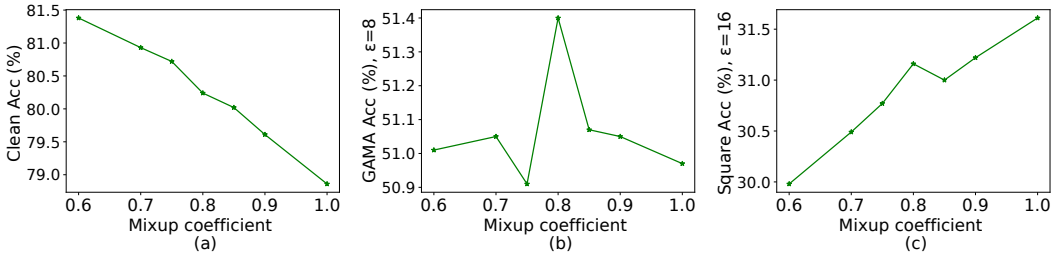


Figure 12: **Sensitivity against variation in Mixup coefficient:** (a) Clean Accuracy (%), (b) Accuracy (%) against GAMA-PGD 100-step attack (Sriramanan et al., 2020) at $\varepsilon = 8/255$ and (c) Accuracy (%) against Square Attack (Andriushchenko et al., 2020) at $\varepsilon = 16/255$ are reported on the CIFAR-10 dataset. The optimal setting chosen is mixup coefficient of 0.8.

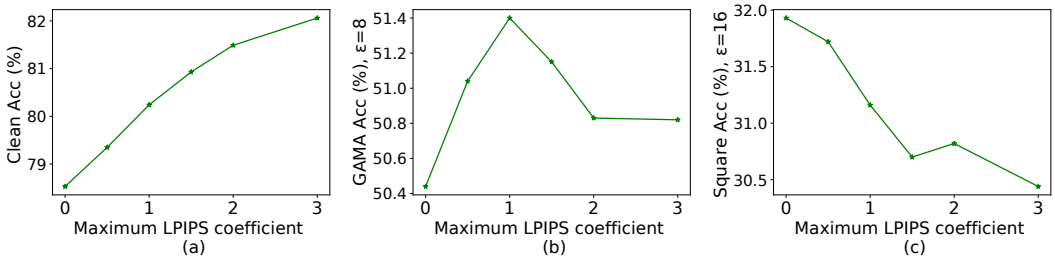


Figure 13: **Sensitivity against variation in Maximum LPIPS coefficient:** (a) Clean Accuracy (%), (b) Accuracy (%) against GAMA-PGD 100-step attack (Sriramanan et al., 2020) at $\varepsilon = 8/255$ and (c) Accuracy (%) against Square Attack (Andriushchenko et al., 2020) at $\varepsilon = 16/255$ are reported on the CIFAR-10 dataset. The optimal setting chosen is maximum lpips coefficient of 1.

examples. While a higher 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 coefficeint in Fig.13. 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 ε used in the mixup iteration. We find that a high epsilon perturbations for mixup iterartion leads to weak attack since we project every perturbation to a lower epsilon value while training, resulting in a higher clean accuracy and lower robust accuracy. Overall, we observe that OA-AT is less sensitive to hyperparameter changes.

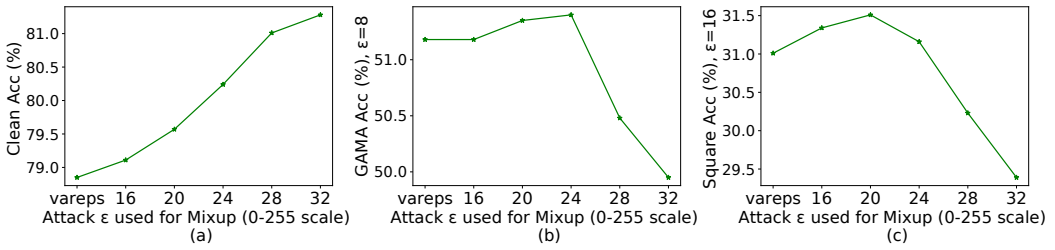


Figure 14: **Sensitivity against variation in ε used in mixup iteration:** (a) Clean Accuracy (%), (b) Accuracy (%) against GAMA-PGD 100-step attack (Sriramanan et al., 2020) at $\varepsilon = 8/255$ and (c) Accuracy (%) against Square Attack (Andriushchenko et al., 2020) at $\varepsilon = 16/255$ are reported on the CIFAR-10 dataset. The optimal setting chosen is $\varepsilon = 24$ for mixup.

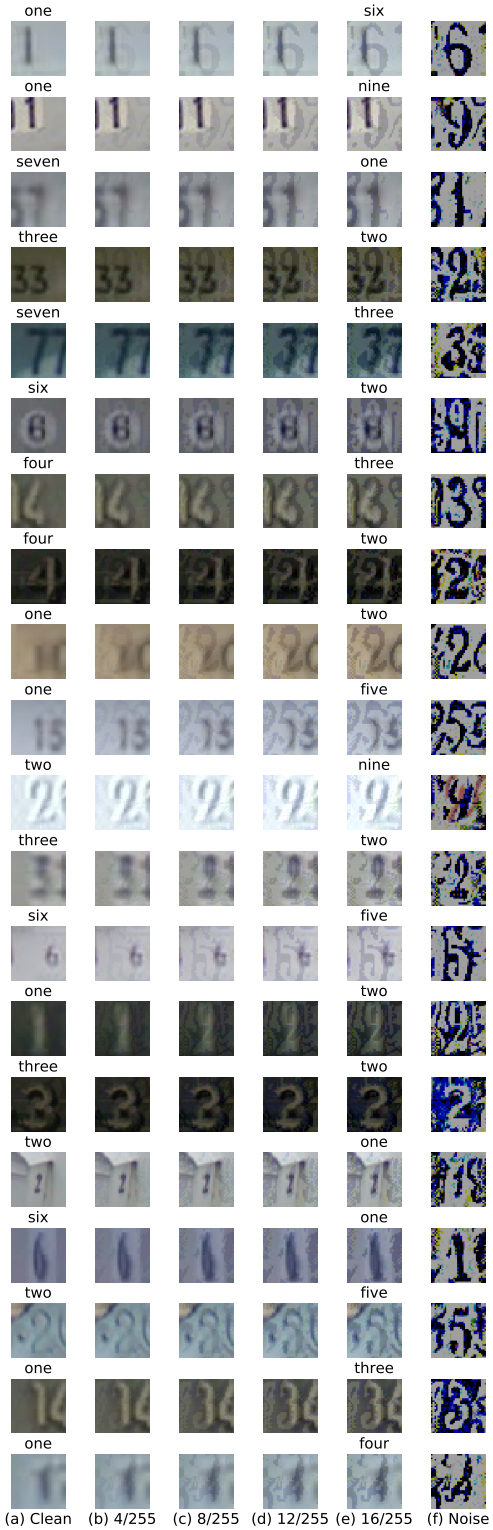


Figure 15: **SVHN, Low-Contrast**

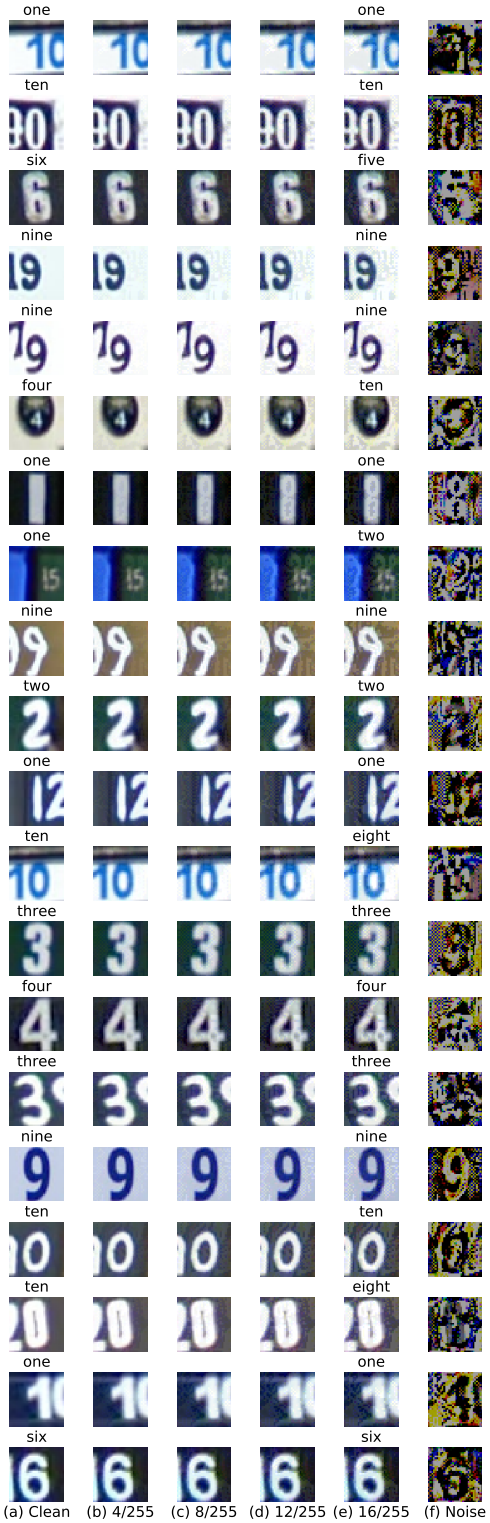


Figure 16: **SVHN, High Contrast**

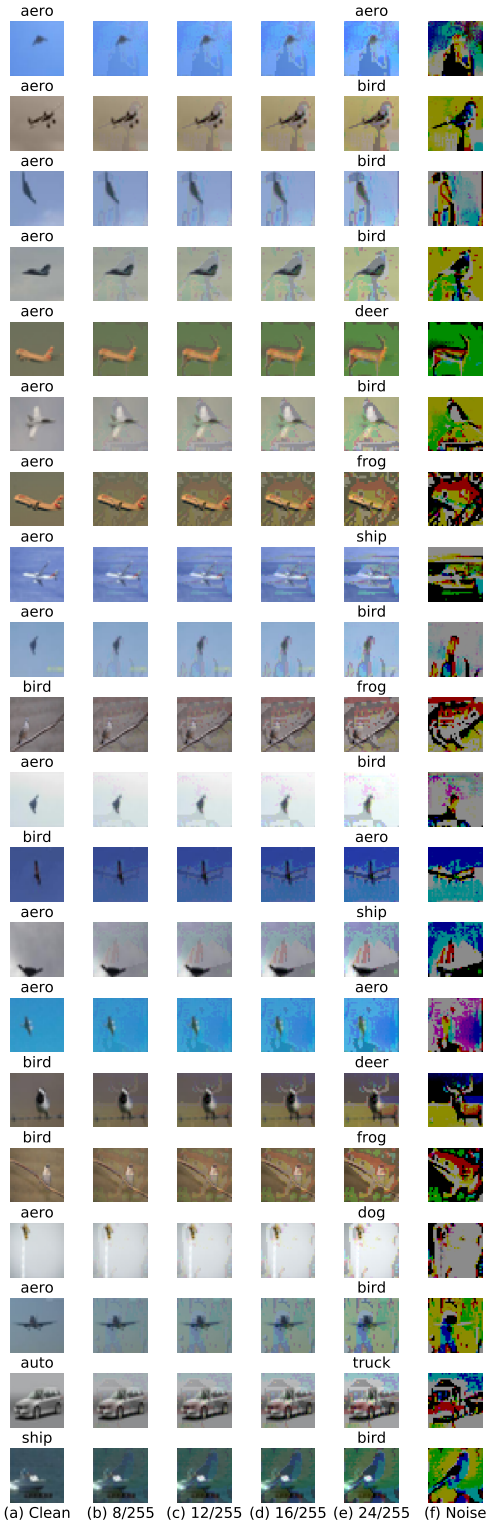


Figure 17: **CIFAR-10 Low Contrast**

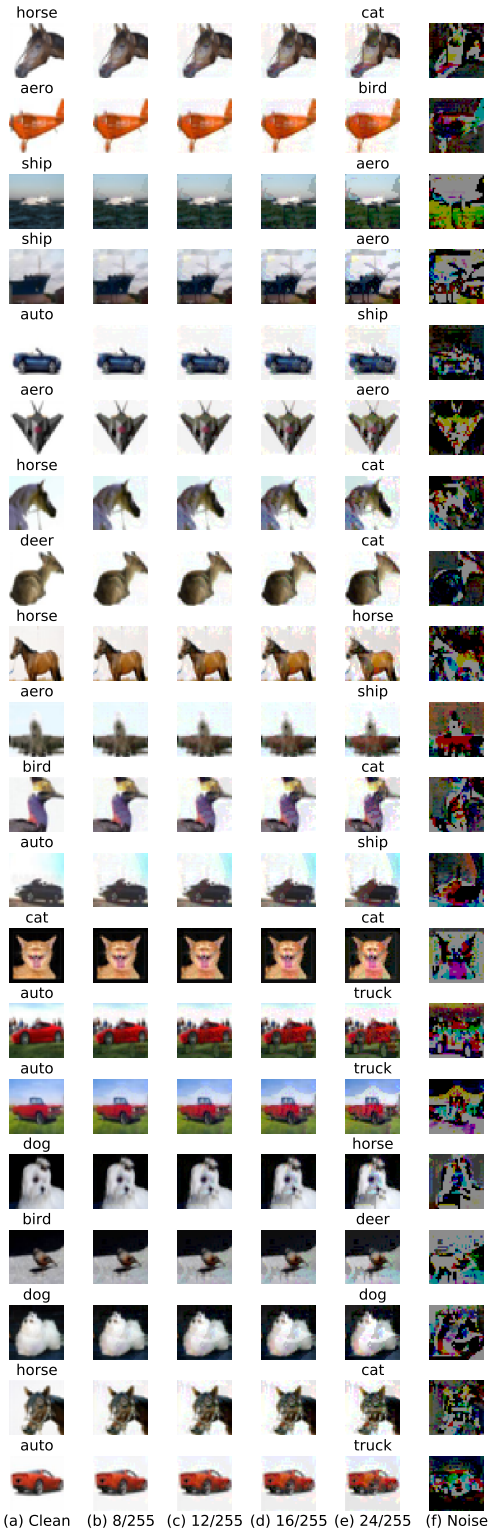


Figure 18: **CIFAR-10 High Contrast**

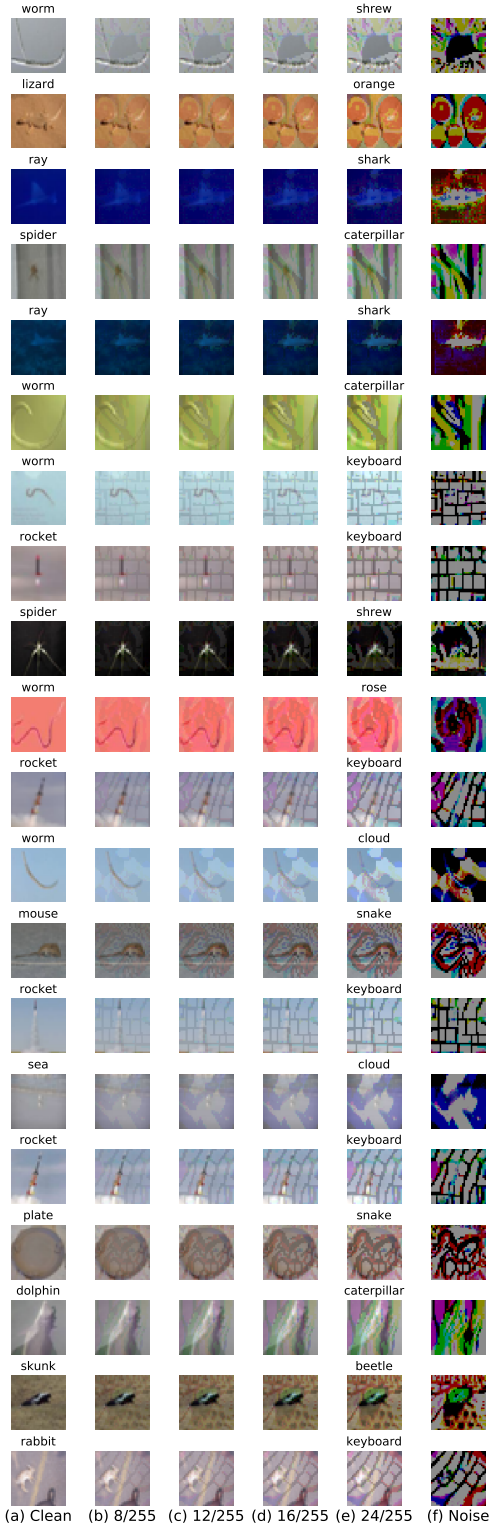


Figure 19: **CIFAR-100 Low Contrast**

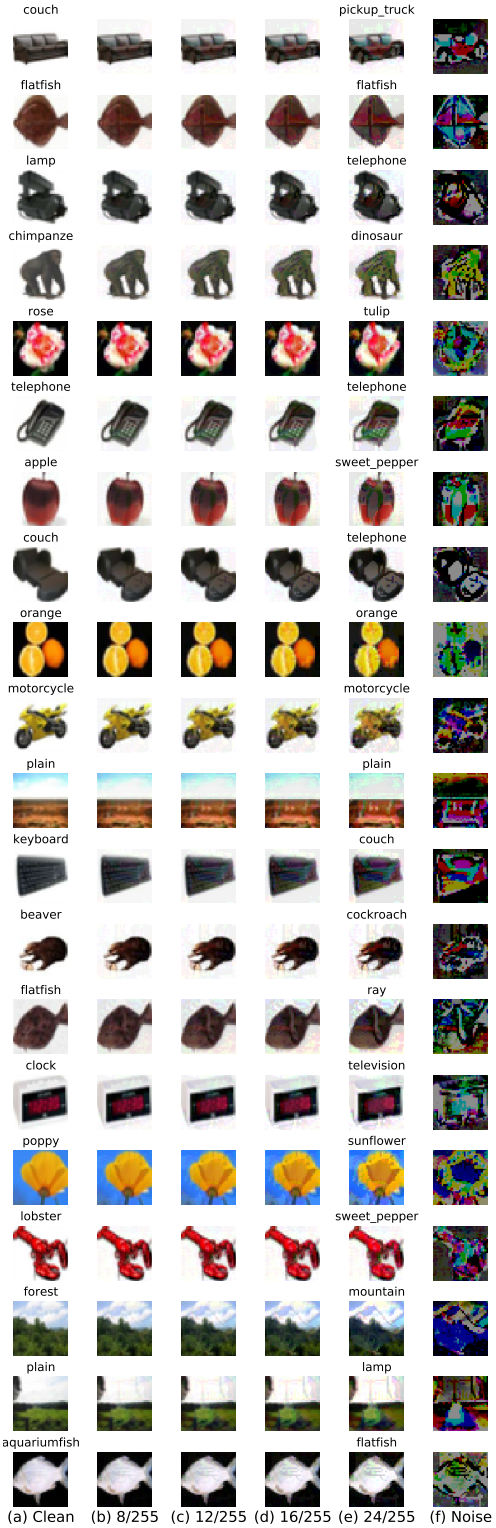


Figure 20: **CIFAR-100 High Contrast**