Supplementary Material: Towards Achieving Adversarial Robustness Beyond Perceptual Limits

Sravanti Addepalli*, Samyak Jain*, Gaurang Sriramanan, Shivangi Khare, R.Venkatesh Babu
Video Analytics Lab, Indian Institute of Science, Bangalore, India

A1. Oracle-Invariant Attacks

Square Attack: The strongest Oracle-Invariant examples are generated using the Square attack [2]. As shown in Fig.A1, these are Oracle-Invariant since this is a query-based attack and does not use gradients from any model for attack generation. However this attack uses 5000 queries, and is a computationally expensive attack. Hence this attack cannot be used for adversarial training, although it is one of the best attacks for evaluations. We note that reducing the number of queries makes it computationally efficient, however it also reduces the effectiveness of the attack significantly.

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:

$$\mathcal{L}_{CE}(x, y) - \lambda_{Disc} \cdot \mathcal{L}_{BCE}(\hat{x}, \text{OI}) \tag{A1}$$

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.A1 therefore attempts to break the most salient property of Oracle-Sensitive attacks, which is the spatial correlation between an image and its perturbation.

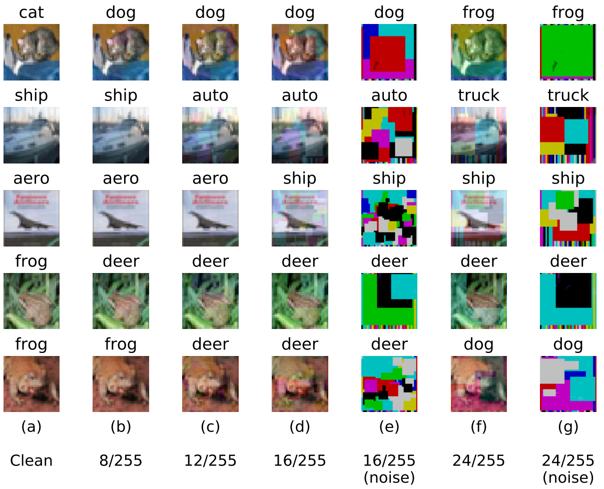
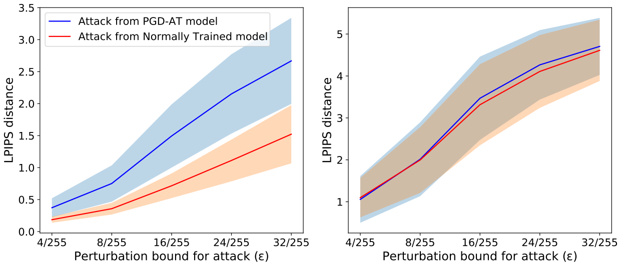


Figure A1. **Square attack:** Adversarially attacked images (b, c, d, f) and the corresponding perturbations (e, g) for various ℓ_∞ bounds generated using the gradient-free random search based attack Square [2]. The clean image is shown in (a). Attacks are generated from a model trained using the proposed Oracle-Aligned Adversarial Training (OA-AT) algorithm on CIFAR-10. Prediction of the same model is printed above each image.

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 [24, 13]. As shown in Fig.A2, while the standard AlexNet model that is used in prior work [13] 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 at-

*Equal contribution.

Correspondence to: Sravanti Addepalli, sravantia@iisc.ac.in



(a) PGD-AT LPIPS model

(b) Normally Trained LPIPS model

Figure A2. LPIPS distance between clean and adversarially perturbed images. Attacks generated from PGD-AT [14, 15] 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.

tacks generated from a PGD-AT [14, 15] model (Fig.1(c-e) of the main paper) as Oracle-Sensitive attacks, and attacks generated from a Normally Trained model (Fig.1(h) of 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:

$$\mathcal{L}_{CE}(x, y) - \lambda_{\text{LPIPS}} \cdot \text{LPIPS}(x, \hat{x}) \tag{A2}$$

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.A3, 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.A3(a). This can be observed more distinctly at perturbation bounds of $24/255$ and $32/255$. The perturbations in Fig.A3(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.A3(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-A1, 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 [22] against the two proposed Oracle-Invariant attacks, LPIPS based attack and Discriminator based attack in Fig.A4 (a) and (b) respectively. We vary the coefficient of the regularizers used in the generation of attacks, λ_{Disc} (Eq.A1) and λ_{LPIPS} (Eq.A2) 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 [22] baseline.

A2. Details on the Datasets used

We evaluate the proposed approach on the CIFAR-10 and CIFAR-100 [12] datasets. The two datasets consist of RGB images of spatial dimension 32×32, and contain 10 and 100 distinct classes respectively. 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. 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. 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 both 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. Further, we consider the ℓ_∞ threat model of radius $16/255$ to investigate robustness within moderate magnitude perturbation bounds.

A3. Details on Training

The algorithm for the proposed method as explained in Sec.4 (main paper) is presented in Algorithm-A1. 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-A1). For CIFAR-10 training on ResNet-18, we set the weight of the adversarial loss L_{adv} in L21 of Alg.A1 (β parameter of TRADES [23]) 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 increased from $12/255$ to $16/255$. In this moderate perturbation regime, we also linearly increase the coefficient of the LPIPS distance (Alg.A1, L17) from 0 to 1, and linearly decrease the α parameter used in the convex combination of softmax prediction (Alg.A1, L14) from 1 to 0.8. This results in a smooth transition from adversarial training on imperceptible attacks to attacks with larger perturbation

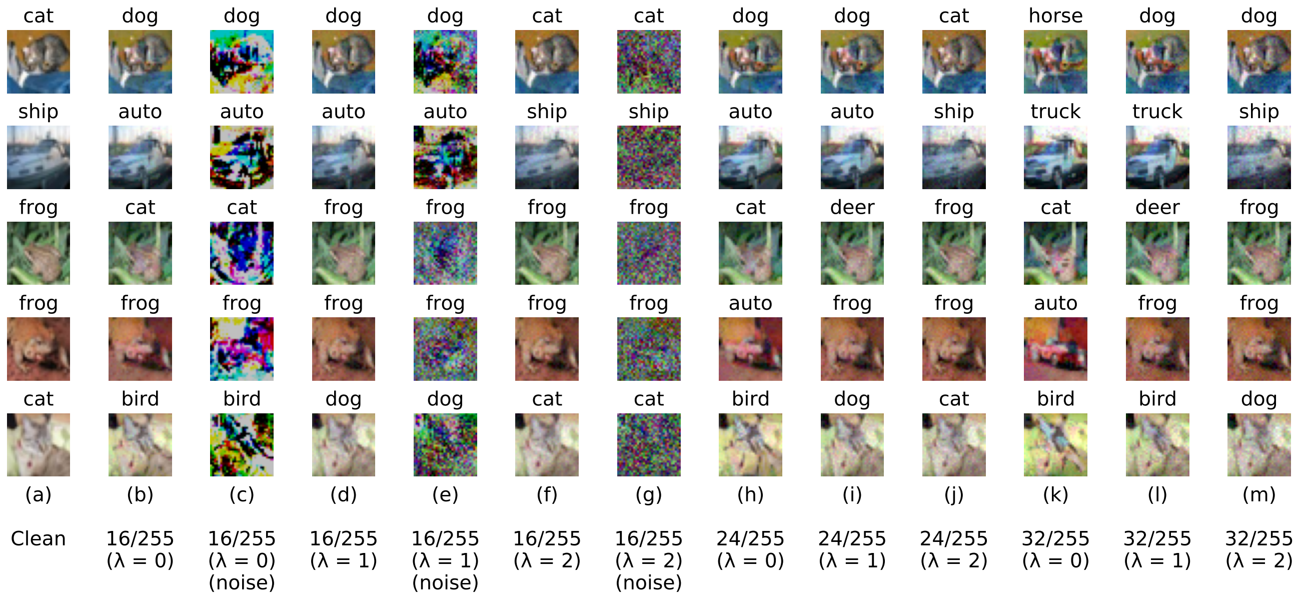

cat dog dog dog dog cat cat dog dog cat horse dog dog
ship auto auto auto auto ship ship auto auto ship truck truck ship
frog cat cat frog frog frog frog cat deer frog cat deer frog
frog frog frog frog frog frog frog auto frog frog auto frog frog
cat bird bird dog dog cat cat bird dog cat bird bird dog
(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m)
Clean
16/255 (λ = 0)
16/255 (λ = 0) (noise)
16/255 (λ = 1)
16/255 (λ = 1) (noise)
16/255 (λ = 2)
16/255 (λ = 2) (noise)
24/255 (λ = 0)
24/255 (λ = 1)
24/255 (λ = 2)
32/255 (λ = 0)
32/255 (λ = 1)
32/255 (λ = 2)

Figure A3. Oracle-Invariant adversarial examples generated using the LPIPS based PGD attack in Eq.A2 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.

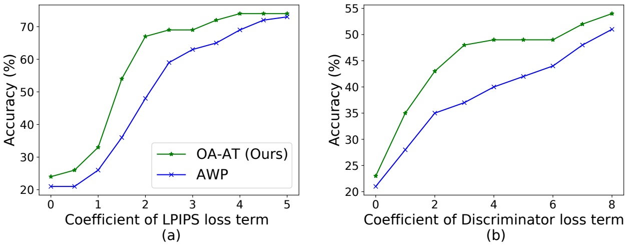


Figure A4. Comparison of the proposed model with the best baseline, AWP [22] trained on CIFAR-10 with ResNet-18 architecture, against attacks of varying strength and Oracle sensitivity constrained within perturbation bound of $\varepsilon = 16/255$. (a) LPIPS based regularizer, and (b) Discriminator 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.

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. We compute the LPIPS distance using an exponential weight averaged model with $\tau = 0.995$. We note from Table-A1 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 [7] for training on CIFAR-100, 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-A1), 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-A1) to 1 (E12 in Table-A1).

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.

Table A1. **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.

	CIFAR-10				**CIFAR-100**			
Method	**Clean**	**GAMA (8/255)**	**GAMA (16/255)**	**Square (16/255)**	**Clean**	**GAMA (8/255)**	**GAMA (16/255)**	**Square (16/255)**
E1: OAAT (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

Table A2. **CIFAR-10**: Performance (%) of the proposed defense OA-AT against attacks with different ε bounds, when compared to the following baselines: AWP [22], ExAT [16], TRADES [23], ATES [17], PGD-AT [14] and FAT [1]. AWP [22] 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 [6] (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	$+0.32$	$+1.68$	$+1.34$	$+1.01$	$+0.46$	$+3.07$	$+2.54$	$+2.15$	$+1.30$	$+4.97$	$+3.07$	$+2.65$

A3.1. 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-A1. 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 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-A2.

As discussed in Sec.4 (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 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

Table A3. **CIFAR-100**: Performance (%) of the proposed defense OA-AT against attacks with different ε bounds, when compared to the following baselines: AWP [22], ExAT [16], TRADES [23], ATES [17], PGD-AT [14] and FAT [1]. AWP [22] is the strongest baseline. The baselines are sorted by AutoAttack accuracy [6] (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 [7] as the augmentation strategy, we present results on the strongest baseline AWP [22] 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

Algorithm A1 Oracle-Aligned Adversarial Training

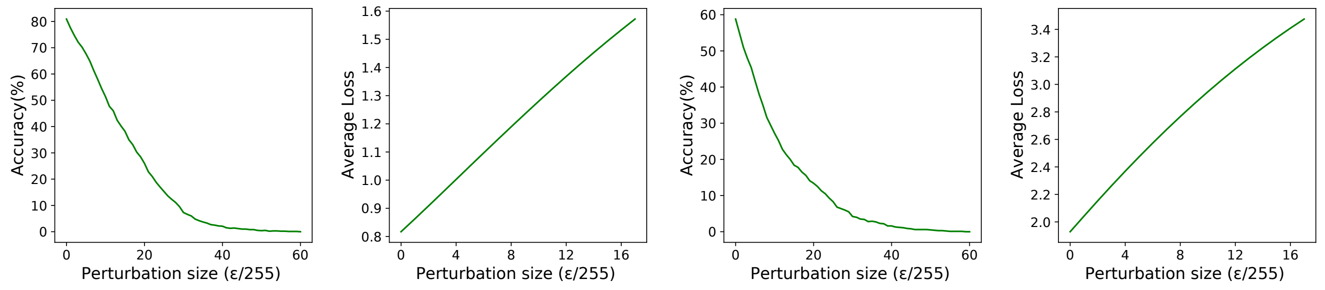
```
Input: Deep Neural Network $f_\theta$ with parameters $\theta$, Training Data $\{x_i, y_i\}_{i=1}^{M}$, Epochs $T$, Learning Rate $\eta$, Perturbation budget $\varepsilon_{max}$, Adversarial Perturbation function $A(x, y, \ell, \varepsilon)$ which maximises loss $\ell$
for epoch $= 1$ to $T$ do
    $\widetilde{\varepsilon} = \max\{\varepsilon_{max}/4, \varepsilon_{max} \cdot \text{epoch}/T\}$
    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
            $\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i)$
            $\widetilde{\delta}_i = A(x_i, y_i, \ell, \widetilde{\varepsilon})$
            $L_{adv} = \text{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) || f_\theta(x_i)\right)$
        else if $i \,\%\, 2 = 0$ then
            $\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i)$
            $\widehat{\delta}_i = A(x_i, y_i, \ell, 1.5 \cdot \varepsilon_{max})$
            $\widetilde{\delta}_i = \Pi_\infty(\widehat{\delta}_i, \widetilde{\varepsilon})$
            $L_{adv} = \text{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) \quad ||$
                        $\alpha \cdot f_\theta(x_i) + (1 - \alpha) \cdot f_\theta(x_i + \widehat{\delta}_i)\right)$
        else
            $\delta_i \sim U(-\widetilde{\varepsilon}, \widetilde{\varepsilon})$
            $\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i) - \text{LPIPS}(x_i, x_i + \delta_i)$
            $\widetilde{\delta}_i = A(x_i, y_i, \ell, \widetilde{\varepsilon})$
            $L_{adv} = \text{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) || f_\theta(x_i)\right)$
        end if
        $L = \ell_{CE}(f_\theta(x_i), y_i) + L_{adv}$
        $\theta = \theta - \eta \cdot \nabla_\theta L$
    end for
end for
```

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*. [9].

A4. Detailed Results

In Tables-A2 and A3, we present results of different defense methods such as AWP-TRADES [22], TRADES [23], PGD-AT [14], ExAT [16], ATES [17] and FAT [1], evaluated across a wide range of adversarial attacks. We present evaluations on the Black-Box FGSM attack [8] 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 [21], the GAMA PGD-100 attack [18] and AutoAttack [6], with the latter two being amongst the strongest of attacks known to date. Lastly, we also present evaluations on the Square attack [2] 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-A2. 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 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-A2. We observe that the proposed method OA-AT consistently outperforms other approaches on all three metrics described in Sec.3.3 (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 ℓ_∞ constraint set of $8/255$ when compared to



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

Figure A5. 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 [3] (b, d) Plots showing the variation of Cross-Entropy Loss on FGSM attack [8] 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 [3]

existing approaches, despite being trained on larger perturbations sets.

CIFAR-100: In Table-A3, 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 [7, 19]. To enable fair comparison, we incorporate AutoAugment for the strongest baseline, AWP [22] 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.

A5. Gradient Masking Checks

As discussed by Athalye *et al.* [3], we present various checks to ensure the absence of Gradient Masking in the proposed defense. In Fig.A5(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.A5(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.

Table A4. **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† [6]	100	20	**50.88**
GAMA-MT [18]	100	5	50.90
ODS (98 +2 steps) [20]	100	100	50.94
MDMT attack [11]	100	10	51.19
Logit-Scaling attack [4, 10]	100	20	51.26
GAMA-PGD [18]	100	1	51.40
MD attack [11]	100	1	51.47
PGD-50 (1000 RR) [14]	50	1000	55.37
PGD-1000 [14]	1000	1	56.15

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-A2,A3. 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-A4, we perform exhaustive evaluations using various attack techniques to further verify the absence of gradient masking. In addition to AutoAttack [6] which in itself consists of an ensemble of four attacks- AutoPGD with Cross-Entropy and Difference-of-Logits loss, the FAB attack [5] and Square Attack [2], we present evaluations against strong multi-targeted attacks such as GAMA-MT [18] and the MDMT attack [11] 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 [20] 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 [4, 10] 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.

References

[1] Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning (ICML)*, 2020.

[2] 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.

[3] 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.

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

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

[6] 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)*, 2020.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

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

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

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

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.