Appendix

A. Differences with ADMM attack [36] and StrAttack [31]

There is a significant technical difference between our proposed attack and the two attacks based on ADMM approaches [36, 31] as the multipliers in the ADMM attacks are used for the problem-splitting constraints, but not for the attack constraints as in our ALM. For two terms, ADMM replaces a one-variable problem of the form $\min_x f(x) + g(x)$ by a two-variable problem:

$$\min_{x,y} f(x) + g(y) \quad \text{s.t.} \quad x = y \tag{11}$$

a splitting that gives raise to variable-consistency constraints $x = y$. In fact, both ADMM attacks [31, 36] are based on a decomposition of the Carlini-Wagner penalty formulation (Equation 7 in [36] and Equation 4 in [31]):

$$D(x + \delta, x) + g_{\text{CW}}(\beta) \quad \text{s.t.} \quad \delta = \beta \tag{12}$$

where D is the distance and g_{CW} is the standard CW penalty for attack constraint $f_y(\boldsymbol{x} + \delta) - \max_{k \neq y} f_k(\boldsymbol{x} + \delta) < 0$; see Equation 5 in [36] and Equation 3 in [31]. The Lagrange multipliers in these ADMM attacks are for the decomposition constraints $\delta = \beta$, but the attack constraints are still handled with the standard CW penalty. In our case, we address the attack constraints with augmented Lagrangian principles, and there is no ADMM splitting in our method. The ADMM attack in [36] has no public implementation, so we were not able to implement it in our experimental framework. The publicly available implementation of the StrAttack [31] contains several differences with the original paper regarding hyper-parameters and update rules for the auxiliary variables. Therefore, we contacted the authors of both papers (of which several are in common) regarding the lack of public implementation, and discrepancies between paper and code, but did not get any answer. Therefore, we did not include these attacks in our experiments. It should be noted that StrAttack's [31] implementation is based on the C&W ℓ_2 attack, but adds a sparsity objective, which tends to increase the perturbation size in terms of ℓ_2 norm compared to the vanilla C&W ℓ_2 attack.

B. CIEDE2000

The CIEDE2000 color difference formula is complex, so we advise the reader to look at the original work [27]. This formula is calculated using the CIELAB color space. However, most image datasets are provided in an RGB format. Therefore, to use the CIEDE2000 color difference formula, we must first convert the images from RGB to the CIELAB color space. We need to use a first conversion between RGB and XYZ color spaces. For this step, we need to know the RGB working space and the reference white. However, we do not have that information available, as we do not know how the images were captured in the first place. As a consequence, we make the assumption of the sRGB working space with an Illuminant D65 white reference. With these assumptions, the formula to convert from RGB (with values in $[0, 1]$) to XYZ is the following:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124564 & 0.3575761 & 0.1804375 \\ 0.2126729 & 0.7151522 & 0.0721750 \\ 0.0193339 & 0.1191920 & 0.9503041 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{13}$$

Once we have the colors represented in the XYZ color space, we need to convert to the CIELAB color space. The conversion is the following:

$$\begin{aligned} L^\star &= 116 f\left(\frac{Y}{Y_n}\right) - 16 \\ a^\star &= 500\left(f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right) \\ b^\star &= 200\left(f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right) \end{aligned} \tag{14}$$

where:

$$f(t) = \begin{cases} \sqrt[3]{t} & \text{if} \quad t > \delta^3 \\ \frac{t}{3\delta^2} + \frac{4}{29} & \text{otherwise} \end{cases} \tag{15}$$

with $\delta = \frac{6}{29}$. Under the Illuminant D65 white reference, we have $X_n = 95.0489$, $Y_n = 100$ and $Z_n = 108.8840$.

C. Modified DLR loss

The original DLR loss proposed in [13] is formulated as follows:

$$\text{DLR}(\boldsymbol{z}, y) = -\frac{\boldsymbol{z}_y - \max\limits_{i \neq y} \boldsymbol{z}_i}{\boldsymbol{z}_{\pi_1} - \boldsymbol{z}_{\pi_3}} \tag{16}$$

where $\boldsymbol{z} = f(\boldsymbol{x})$ and π is the ordering of the element of $\boldsymbol{z}$ in decreasing order. If a sample $\boldsymbol{x}$ is correctly classified, we have $\text{DLR}(\boldsymbol{z}, y) \in [-1, 0]$ and $\boldsymbol{x}$ is misclassified only if $\text{DLR}(\boldsymbol{z}, y) > 0$. Croce *et al.* also propose a variant for untargeted attacks with a targeted objective. In some cases, performing a targeted attack against each class proved to be more successful at finding untargeted adversarial examples than simply performing an untargeted attack. The targeted variant is:

$$\text{tDLR}(\boldsymbol{z}, y) = -\frac{\boldsymbol{z}_y - \boldsymbol{z}_t}{\boldsymbol{z}_{\pi_1} - (\boldsymbol{z}_{\pi_3} + \boldsymbol{z}_{\pi_4})/2} \tag{17}$$

where t is the target class. For this variant, we can have no guarantee as to what $\boldsymbol{x}$ is classified as, simply by looking at the value of tDLR.

For our optimization problem, we need to have a loss that is negative only when the misclassification or targeted classification is achieved. This way, we can formulate the misclassification or targeted classification constraint as $g(x) < 0$.

To this end, we modify DLR by taking the negative as follows:

$$\mathrm{DLR}^+(\boldsymbol{z}, y) = \frac{\boldsymbol{z}_y - \max_{i \neq y} \boldsymbol{z}_i}{\boldsymbol{z}_{\pi_1} - \boldsymbol{z}_{\pi_3}} \tag{4}$$

For targeted attack, we modify tDLR as follows:

$$\mathrm{tDLR}^+(\boldsymbol{z}, y) = \frac{\max_{i \neq t} \boldsymbol{z}_i - \boldsymbol{z}_t}{\boldsymbol{z}_{\pi_1} - (\boldsymbol{z}_{\pi_3} + \boldsymbol{z}_{\pi_4})/2} \tag{18}$$

With these modifications, the misclassification and targeted classification constraints are respected only when DLR^+ and tDLR^+ are negative. Conversely, the constraints are violated when these losses are positive, hence the $^+$ superscript.

D. Penalty functions

The four penalty functions plotted in Figure 1 are taken from [4] and defined as follows:

$$\mathrm{PHR}(y, \rho, \mu) = \frac{1}{2\rho}(\max\{0, \mu + \rho y\}^2 - \mu^2) \tag{19}$$

$$P_1(y, \rho, \mu) = \begin{cases} \mu y + \frac{1}{2}\rho y^2 + \rho^2 y^3 & \text{if} \quad y \geq 0 \\ \mu y + \frac{1}{2}\rho y^2 & \text{if} \quad -\frac{\mu}{\rho} \leq y \leq 0 \\ -\frac{1}{2\rho}\mu^2 & \text{if} \quad y \leq -\frac{\mu}{\rho} \end{cases} \tag{20}$$

$$P_2(y, \rho, \mu) = \begin{cases} \mu y + \mu\rho y^2 + \frac{1}{6}\rho^2 y^3 & \text{if} \quad y \geq 0 \\ \frac{\mu y}{1 - \rho y} & \text{if} \quad y \leq 0 \end{cases} \tag{21}$$

$$P_3(y, \rho, \mu) = \begin{cases} \mu y + \mu\rho y^2 & \text{if} \quad y \geq 0 \\ \frac{\mu y}{1 - \rho y} & \text{if} \quad y \leq 0 \end{cases} \tag{22}$$

E. Choice of α

For the experiments, we change the value of α according to the number of iterations. In our attack, α is a smoothing parameter. However, too much smoothing can degrade the performance of the attack for lower numbers of iterations. Therefore, we recommend values between 0.5 and 0.9 for numbers of iterations between 100 and 1 000, and to keep $\alpha = 0.9$ for more than 1 000 iterations. Since our attack aims to find minimal adversarial perturbations, lower numbers of iterations are not recommended.

F. Hyper-parameter ϵ values

Table 3 reports the different ϵ used for each distance function in our experiments.

G. Additional results with SSIM

We also tested our ALMA attack with the SSIM [29] for which, to the best of our knowledge, no gradient-based attack currently exists. The only related work on adversarial attacks and SSIM is a black-box method [16]. SSIM is a similarity function, so identical images have a SSIM of 1. Therefore, we minimize the quantity $1 - \mathrm{SSIM}$ and report this instead of the SSIM. The SSIM metric is defined between two gray-level images. Several modifications exist for color images, however, we simply considered the average SSIM over the color channels. Tables 10 and 15 report the results for all models on CIFAR10 and ImageNet respectively.

Distance	ϵ
ℓ_1	0.5
ℓ_2	0.1
SSIM	3×10^{-5}
CIEDE2000	0.05
LPIPS	1×10^{-3}

Table 3: Initial values of ϵ for each distance.

H. Detailed experimental results

Tables 4, 5, 6, 7, 10, 8, 9, 11, 12, 15, 13 and 14 report the detailed results for each dataset, model and attack. Results from Tables 1 and 2 are calculated from these tables using the geometric mean over the models. For the CIFAR10 and ImageNet models, RN stands for ResNet and WRN for Wide ResNet. For ImageNet, the targeted variant of FAB is used in the experiments denoted by a T superscript (see Section 4 for details).

I. Robust Accuracy curves

Figures 6, 7, 8, 9, 10, 12, 13 and 14 present the robust accuracy curves for each dataset and model against the attacks considered for each distance. The dotted line represent the reduced budget versions of the attack, as reported in the corresponding tables.

Model	Attack	ASR (%)	Median ℓ_1	forwards / backwards
SmallCNN	EAD 9×100 [9]	100	8.41	870 / 439
	EAD 9×1000 [9]	100	7.90	3 810 / 1 909
	FAB ℓ_1 100 [12]	100	6.31	201 / 1 000
	FAB ℓ_1 1000 [12]	100	6.20	2 001 / 10 000
	FMN ℓ_1 100 [25]	95.02	6.50	100 / 100
	FMN ℓ_1 1000 [25]	95.19	6.39	1 000 / 1 000
	ALMA ℓ_1 100	100	6.77	100 / 100
	ALMA ℓ_1 1000	100	6.23	1 000 / 1 000
SmallCNN DDN	EAD 9×100 [9]	100	17.41	990 / 499
	EAD 9×1000 [9]	100	16.35	5 010 / 2 509
	FAB ℓ_1 100 [12]	100	16.53	201 / 1 000
	FAB ℓ_1 1000 [12]	100	15.42	2 001 / 10 000
	FMN ℓ_1 100 [25]	99.97	16.09	100 / 100
	FMN ℓ_1 1000 [25]	99.97	15.67	1 000 / 1 000
	ALMA ℓ_1 100	100	14.73	100 / 100
	ALMA ℓ_1 1000	100	14.02	1 000 / 1 000
SmallCNN TRADES	EAD 9×100 [9]	100	14.80	967 / 486
	EAD 9×1000 [9]	100	12.15	6 409 / 3 208
	FAB ℓ_1 100 [12]	99.22	36.60	201 / 1 000
	FAB ℓ_1 1000 [12]	99.35	32.37	2 001 / 10 000
	FMN ℓ_1 100 [25]	50.30	42.02	100 / 100
	FMN ℓ_1 1000 [25]	98.30	8.28	1 000 / 1 000
	ALMA ℓ_1 100	100	6.16	100 / 100
	ALMA ℓ_1 1000	100	5.32	1 000 / 1 000
CROWN IBP	EAD 9×100 [9]	89.17	106.95	509 / 258
	EAD 9×1000 [9]	91.32	86.45	5 210 / 2 609
	FAB ℓ_1 100 [12]	99.99	147.79	201 / 1 000
	FAB ℓ_1 1000 [12]	99.99	110.96	2 001 / 10 000
	FMN ℓ_1 100 [25]	49.89	–	100 / 100
	FMN ℓ_1 1000 [25]	88.36	3.50	1 000 / 1 000
	ALMA ℓ_1 100	99.59	27.94	100 / 100
	ALMA ℓ_1 1000	100	5.65	1 000 / 1 000

Table 4: Performance of the ℓ_1 attacks on the MNIST dataset for each model.

Model	Attack	ASR (%)	Median ℓ_2	forwards / backwards
SmallCNN	C&W ℓ_2 9×1000 [7]	99.98	1.35	9 000 / 9 000
	C&W ℓ_2 9×10 000 [7]	99.77	1.35	90 000 / 90 000
	DDN 100 [26]	100	1.39	100 / 100
	DDN 1000 [26]	100	1.37	1 000 / 1 000
	FAB ℓ_2 100 [12]	100	1.37	201 / 1 000
	FAB ℓ_2 1000 [12]	100	1.36	2 001 / 10 000
	FMN ℓ_2 100 [25]	82.04	1.53	100 / 100
	FMN ℓ_2 1000 [25]	96.61	1.39	1 000 / 1 000
	$\text{APGD}^{\text{T}}_{\text{DLR}}$ ℓ_2 [13]	100	1.31	13 082 / 13 062
	ALMA ℓ_2 100	100	1.38	100 / 100
	ALMA ℓ_2 1000	100	1.32	1 000 / 1 000
SmallCNN DDN	C&W ℓ_2 9×1000 [7]	99.96	2.76	9 000 / 9 000
	C&W ℓ_2 9×10 000 [7]	99.59	2.69	90 000 / 90 000
	DDN 100 [26]	100	2.74	100 / 100
	DDN 1000 [26]	100	2.66	1 000 / 1 000
	FAB ℓ_2 100 [12]	100	2.74	201 / 1 000
	FAB ℓ_2 1000 [12]	100	2.71	2 001 / 10 000
	FMN ℓ_2 100 [25]	99.95	2.67	100 / 100
	FMN ℓ_2 1000 [25]	100	2.67	1 000 / 1 000
	$\text{APGD}^{\text{T}}_{\text{DLR}}$ ℓ_2 [13]	100	2.58	13 410 / 13 390
	ALMA ℓ_2 100	100	2.68	100 / 100
	ALMA ℓ_2 1000	100	2.59	1 000 / 1 000
SmallCNN TRADES	C&W ℓ_2 9×1000 [7]	99.99	3.32	9 000 / 9 000
	C&W ℓ_2 9×10 000 [7]	99.97	2.28	90 000 / 90 000
	DDN 100 [26]	99.69	2.17	100 / 100
	DDN 1000 [26]	100	1.91	1 000 / 1 000
	FAB ℓ_2 100 [12]	99.88	1.77	201 / 1 000
	FAB ℓ_2 1000 [12]	99.90	1.74	2 001 / 10 000
	FMN ℓ_2 100 [25]	86.41	2.24	100 / 100
	FMN ℓ_2 1000 [25]	99.83	1.99	1 000 / 1 000
	$\text{APGD}^{\text{T}}_{\text{DLR}}$ ℓ_2 [13]	100	3.36	13 919 / 13 899
	ALMA ℓ_2 100	100	1.74	100 / 100
	ALMA ℓ_2 1000	100	1.55	1 000 / 1 000
CROWN IBP	C&W ℓ_2 9×1000 [7]	2.61	–	9 000 / 9 000
	C&W ℓ_2 9×10 000 [7]	2.63	–	90 000 / 90 000
	DDN 100 [26]	94.34	1.46	100 / 100
	DDN 1000 [26]	99.27	0.97	1 000 / 1 000
	FAB ℓ_2 100 [12]	99.98	5.19	201 / 1 000
	FAB ℓ_2 1000 [12]	99.98	3.34	2 001 / 10 000
	FMN ℓ_2 100 [25]	67.80	2.14	100 / 100
	FMN ℓ_2 1000 [25]	89.08	1.34	1 000 / 1 000
	$\text{APGD}^{\text{T}}_{\text{DLR}}$ ℓ_2 [13]	99.94	3.57	9 286 / 9 273
	ALMA ℓ_2 100	98.90	4.96	100 / 100
	ALMA ℓ_2 1000	100	1.26	1 000 / 1 000

Table 5: Performance of the ℓ_2 attacks on the MNIST dataset for each model.

Model	Attack	ASR (%)	Median ℓ_1	forwards / backwards
WRN 28-10	EAD 9×100 [9]	100	1.79	530 / 269
	EAD 9×1000 [9]	100	1.62	4 910 / 2 459
	FAB ℓ_1 100 [12]	92.3	1.27	200 / 1 000
	FAB ℓ_1 1000 [12]	98.8	1.07	2 000 / 10 000
	FMN ℓ_1 100 [25]	99.7	1.01	100 / 100
	FMN ℓ_1 1000 [25]	99.5	0.98	1 000 / 1 000
	ALMA ℓ_1 100	100	1.26	100 / 100
	ALMA ℓ_1 1000	100	1.02	1 000 / 1 000
WRN 28-10 Carmon *et al.* [8]	EAD 9×100 [9]	100	6.62	600 / 304
	EAD 9×1000 [9]	100	6.07	3 760 / 1 884
	FAB ℓ_1 100 [12]	97.8	5.57	200 / 1 000
	FAB ℓ_1 1000 [12]	98.2	5.07	2 000 / 10 000
	FMN ℓ_1 100 [25]	100	4.70	100 / 100
	FMN ℓ_1 1000 [25]	100	4.64	1 000 / 1 000
	ALMA ℓ_1 100	100	5.20	100 / 100
	ALMA ℓ_1 1000	100	4.75	1 000 / 1 000
RN-50 Augustin *et al.* [1]	EAD 9×100 [9]	100	19.18	590 / 299
	EAD 9×1000 [9]	100	16.39	4 260 / 2 134
	FAB ℓ_1 100 [12]	99.8	10.95	200 / 1 000
	FAB ℓ_1 1000 [12]	99.8	10.09	2 000 / 10 000
	FMN ℓ_1 100 [25]	100	10.21	100 / 100
	FMN ℓ_1 1000 [25]	100	9.79	1 000 / 1 000
	ALMA ℓ_1 100	100	12.15	100 / 100
	ALMA ℓ_1 1000	100	10.35	1 000 / 1 000

Table 6: Performance of the ℓ_1 attacks on the CIFAR10 dataset for each model.

Model	Attack	ASR (%)	Median ℓ_2	forwards / backwards
WRN 28-10	C&W ℓ_2 9×1000 [7]	100	0.10	9 000 / 9 000
	C&W ℓ_2 9×10 000 [7]	100	0.10	90 000 / 90 000
	DDN 100 [26]	100	0.11	100 / 100
	DDN 1000 [26]	100	0.11	1 000 / 1 000
	FAB ℓ_2 100 [12]	100	0.09	201 / 1 000
	FAB ℓ_2 1000 [12]	100	0.09	2 001 / 10 000
	FMN ℓ_2 100 [25]	99.7	0.12	100 / 100
	FMN ℓ_2 1000 [25]	99.5	0.09	1 000 / 1 000
	$\text{APGD}^{\text{T}}_{\text{DLR}}$ ℓ_2 [13]	100	0.09	4 336 / 4 312
	ALMA ℓ_2 100	100	0.09	100 / 100
	ALMA ℓ_2 1000	100	0.09	1 000 / 1 000
WRN 28-10 Carmon *et al.* [8]	C&W ℓ_2 9×1000 [7]	100	0.70	7 502 / 7 500
	C&W ℓ_2 9×10 000 [7]	100	0.70	71 602 / 71 600
	DDN 100 [26]	100	0.72	100 / 100
	DDN 1000 [26]	100	0.71	1 000 / 1 000
	FAB ℓ_2 100 [12]	100	0.71	201 / 1 000
	FAB ℓ_2 1000 [12]	100	0.71	2 001 / 10 000
	FMN ℓ_2 100 [25]	100	0.69	100 / 100
	FMN ℓ_2 1000 [25]	100	0.70	1 000 / 1 000
	$\text{APGD}^{\text{T}}_{\text{DLR}}$ ℓ_2 [13]	100	0.68	5 683 / 5 659
	ALMA ℓ_2 100	100	0.70	100 / 100
	ALMA ℓ_2 1000	100	0.67	1 000 / 1 000
RN-50 Augustin *et al.* [1]	C&W ℓ_2 9×1000 [7]	100	0.96	7 515 / 7 513
	C&W ℓ_2 9×10 000 [7]	100	0.95	73 869 / 73 867
	DDN 100 [26]	100	0.97	100 / 100
	DDN 1000 [26]	100	0.96	1 000 / 1 000
	FAB ℓ_2 100 [12]	100	1.01	201 / 1 000
	FAB ℓ_2 1000 [12]	100	1.00	2 001 / 10 000
	FMN ℓ_2 100 [25]	100	0.95	100 / 100
	FMN ℓ_2 1000 [25]	100	0.96	1 000 / 1 000
	$\text{APGD}^{\text{T}}_{\text{DLR}}$ ℓ_2 [13]	100	0.91	6 198 / 6 174
	ALMA ℓ_2 100	100	0.98	100 / 100
	ALMA ℓ_2 1000	100	0.92	1 000 / 1 000

Table 7: Performance of the ℓ_2 attacks on the CIFAR10 dataset for each model.

Model	Attack	ASR (%)	Median CIEDE2000	forwards / backwards
WRN 28-10	C&W CIEDE2000 9×1000	100	0.23	7 741 / 7 740
	Perc-AL 100 [37]	100	0.86	201 / 100
	Perc-AL 1000 [37]	100	0.72	2 001 / 1 000
	ALMA CIEDE2000 100	100	0.18	100 / 100
	ALMA CIEDE2000 1000	100	0.14	1 000 / 1 000
WRN 28-10 Carmon *et al.* [8]	C&W CIEDE2000 9×1000	100	2.12	6 243 / 6 240
	Perc-AL 100 [37]	100	5.69	201 / 100
	Perc-AL 1000 [37]	100	5.82	2 001 / 1 000
	ALMA CIEDE2000 100	100	3.65	100 / 100
	ALMA CIEDE2000 1000	100	2.08	1 000 / 1 000
RN-50 Augustin *et al.* [1]	C&W CIEDE2000 9×1000	100	1.63	6 303 / 6 300
	Perc-AL 100 [37]	100	4.85	201 / 100
	Perc-AL 1000 [37]	100	4.83	2 001 / 1 000
	ALMA CIEDE2000 100	99.8	1.94	100 / 100
	ALMA CIEDE2000 1000	99.8	1.58	1 000 / 1 000

Table 8: Performance of the CIEDE2000 attacks on the CIFAR10 dataset for each model.

Model	Attack	ASR (%)	Median LPIPS $\times 10^{-2}$	forwards / backwards
WRN 28-10	C&W LPIPS 9×1000	100	0.32	4 565 / 4 560
	LPA‡ [21]	100	4.81	1 129 / 1 119
	ALMA LPIPS 100	100	0.29	100 / 100
	ALMA LPIPS 1000	100	0.12	1 000 / 1 000
WRN 28-10 Carmon *et al.* [8]	C&W LPIPS 9×1000	100	0.50	7 981 / 7 980
	LPA‡ [21]	100	5.06	1 092 / 1 082
	ALMA LPIPS 100	100	6.76	100 / 100
	ALMA LPIPS 1000	100	1.01	1 000 / 1 000
RN-50 Augustin *et al.* [1]	C&W LPIPS 9×1000	100	0.64	8 101 / 8 100
	LPA‡ [21]	100	6.42	1 133 / 1 123
	ALMA LPIPS 100	99.9	7.66	100 / 100
	ALMA LPIPS 1000	100	1.82	1 000 / 1 000

Table 9: Performance of the LPIPS variant of ALMA on the CIFAR10 dataset for each model. ‡A binary search is performed on each sample to get a minimal perturbation attack (Equation 2).

Model	Attack	ASR (%)	Median $1-$SSIM $\times 10^{-4}$	forwards / backwards
WRN 28-10	ALMA SSIM 100	100	0.4	100 / 100
	ALMA SSIM 1000	100	0.1	1 000 / 1 000
Wide ResNet 28-10 Carmon *et al.* [8]	ALMA SSIM 100	100	7.5	100 / 100
	ALMA SSIM 1000	100	2.8	1 000 / 1 000
ResNet-50 Augustin *et al.* [1]	ALMA SSIM 100	100	4.1	100 / 100
	ALMA SSIM 1000	100	2.0	1 000 / 1 000

Table 10: Performance of the SSIM variant of ALMA on the CIFAR10 dataset for each model.

Model	Attack	ASR (%)	Median ℓ_1	forwards / backwards
RN-50	EAD 9×100 [9]	100	6.70	437 / 222
	EAD 9×1000 [9]	100	6.08	4 510 / 2 259
	FAB$^{\mathrm{T}}$ ℓ_1 100 [12]	74.4	9.01	1 810 / 900
	FAB$^{\mathrm{T}}$ ℓ_1 1000 [12]	81.0	4.82	18 010 / 9 000
	FMN ℓ_1 100 [25]	95.6	3.72	100 / 100
	FMN ℓ_1 1000 [25]	94.5	3.43	1 000 / 1 000
	ALMA ℓ_1 100	100	8.47	100 / 100
	ALMA ℓ_1 1000	100	4.25	1 000 / 1 000
RN-50 ℓ_2-AT	EAD 9×100 [9]	100	62.21	458 / 233
	EAD 9×1000 [9]	100	55.16	3 450 / 1 729
	FAB$^{\mathrm{T}}$ ℓ_1 100 [12]	98.5	31.33	1 810 / 900
	FAB$^{\mathrm{T}}$ ℓ_1 1000 [12]	93.2	33.58	18 010 / 9 000
	FMN ℓ_1 100 [25]	100	36.68	100 / 100
	FMN ℓ_1 1000 [25]	100	30.52	1 000 / 1 000
	ALMA ℓ_1 100	100	61.37	100 / 100
	ALMA ℓ_1 1000	100	40.41	1 000 / 1 000
RN-50 ℓ_∞-AT	EAD 9×100 [9]	100	6.40	582 / 295
	EAD 9×1000 [9]	100	6.29	3 410 / 1 709
	FAB$^{\mathrm{T}}$ ℓ_1 100 [12]	95.6	4.36	1 810 / 900
	FAB$^{\mathrm{T}}$ ℓ_1 1000 [12]	93.6	4.33	18 010 / 9 000
	FMN ℓ_1 100 [25]	87.8	4.16	100 / 100
	FMN ℓ_1 1000 [25]	87.7	4.16	1 000 / 1 000
	ALMA ℓ_1 100	100	14.90	100 / 100
	ALMA ℓ_1 1000	100	10.33	1 000 / 1 000

Table 11: Performance of the ℓ_1 attacks on the ImageNet dataset for each model.

Model	Attack	ASR (%)	Median ℓ_2	forwards / backwards
RN-50	C&W ℓ_2 9×1000 [7]	100	0.21	8 775 / 8 775
	C&W ℓ_2 9×10 000 [7]	100	0.21	82 668 / 82 667
	DDN 100 [26]	99.8	0.18	100 / 100
	DDN 1000 [26]	99.9	0.17	1 000 / 1 000
	FAB$^{\mathrm{T}}$ ℓ_2 100 [12]	99.3	0.10	1 810 / 900
	FAB$^{\mathrm{T}}$ ℓ_2 1000 [12]	98.0	0.10	18 010 / 9 000
	FMN ℓ_2 100 [25]	98.9	0.12	100 / 100
	FMN ℓ_2 1000 [25]	99.3	0.10	1 000 / 1 000
	APGD$^{\mathrm{T}}_{\mathrm{DLR}}$ ℓ_2 [13]	100	0.09	4 866 4 838
	ALMA ℓ_2 100	100	0.10	100 / 100
	ALMA ℓ_2 1000	100	0.10	1000 / 1000
RN-50 ℓ_2-AT	C&W ℓ_2 9×1000 [7]	99.9	1.17	6 260 / 6 256
	C&W ℓ_2 9×10 000 [7]	99.9	1.17	57 004 / 52 000
	DDN 100 [26]	99.5	1.09	100 / 100
	DDN 1000 [26]	99.7	1.10	1 000 / 1 000
	FAB$^{\mathrm{T}}$ ℓ_2 100 [12]	100	0.81	1 810 / 900
	FAB$^{\mathrm{T}}$ ℓ_2 1000 [12]	99.3	0.81	18 010 / 9 000
	FMN ℓ_2 100 [25]	99.6	0.84	100 / 100
	FMN ℓ_2 1000 [25]	99.9	0.82	1 000 / 1 000
	APGD$^{\mathrm{T}}_{\mathrm{DLR}}$ ℓ_2 [13]	100	0.80	7 005 / 6 977
	ALMA ℓ_2 100	100	0.85	100 / 100
	ALMA ℓ_2 1000	100	0.84	1 000 / 1 000
RN-50 ℓ_∞-AT	C&W ℓ_2 9×1000 [7]	99.6	0.76	6 933 / 6 930
	C&W ℓ_2 9×10 000 [7]	99.6	0.76	65 203 / 65 200
	DDN 100 [26]	99.8	0.67	100 / 100
	DDN 1000 [26]	100	0.66	1 000 / 1 000
	FAB$^{\mathrm{T}}$ ℓ_2 100 [12]	99.8	0.55	1 810 / 900
	FAB$^{\mathrm{T}}$ ℓ_2 1000 [12]	99.4	0.55	18 010 / 9 000
	FMN ℓ_2 100 [25]	99.8	0.57	100 / 100
	FMN ℓ_2 1000 [25]	99.7	0.57	1 000 / 1 000
	APGD$^{\mathrm{T}}_{\mathrm{DLR}}$ ℓ_2 [13]	100	0.54	6 647 / 6 619
	ALMA ℓ_2 100	100	0.62	100 / 100
	ALMA ℓ_2 1000	100	0.54	1 000 / 1 000

Table 12: Performance of the ℓ_2 attacks on the ImageNet dataset for each model.

Model	Attack	ASR (%)	Median CIEDE2000	forwards / backwards
RN-50	C&W CIEDE2000 9×1000	100	0.80	4 505 / 4 500
	Perc-AL 100 [37]	100	1.31	2 01 / 100
	Perc-AL 1000 [37]	100	1.07	2 001 / 1 000
	ALMA CIEDE2000 100	100	0.17	100 / 100
	ALMA CIEDE2000 1000	100	0.13	1 000 / 1 000
RN-50 ℓ_2-AT	C&W CIEDE2000 9×1000	100	1.64	6 303 / 6 300
	Perc-AL 100 [37]	99.9	5.87	201 / 100
	Perc-AL 1000 [37]	99.9	6.07	2 001 / 1 000
	ALMA CIEDE2000 100	100	1.51	100 / 100
	ALMA CIEDE2000 1000	100	1.34	1 000 / 1 000
RN-50 ℓ_∞-AT	C&W CIEDE2000 9×1000	100	2.05	6 303 / 6 300
	Perc-AL 100 [37]	99.8	5.82	201 / 100
	Perc-AL 1000 [37]	99.9	6.12	2 001 / 1 000
	ALMA CIEDE2000 100	100	1.61	100 / 100
	ALMA CIEDE2000 1000	100	1.46	1 000 / 1 000

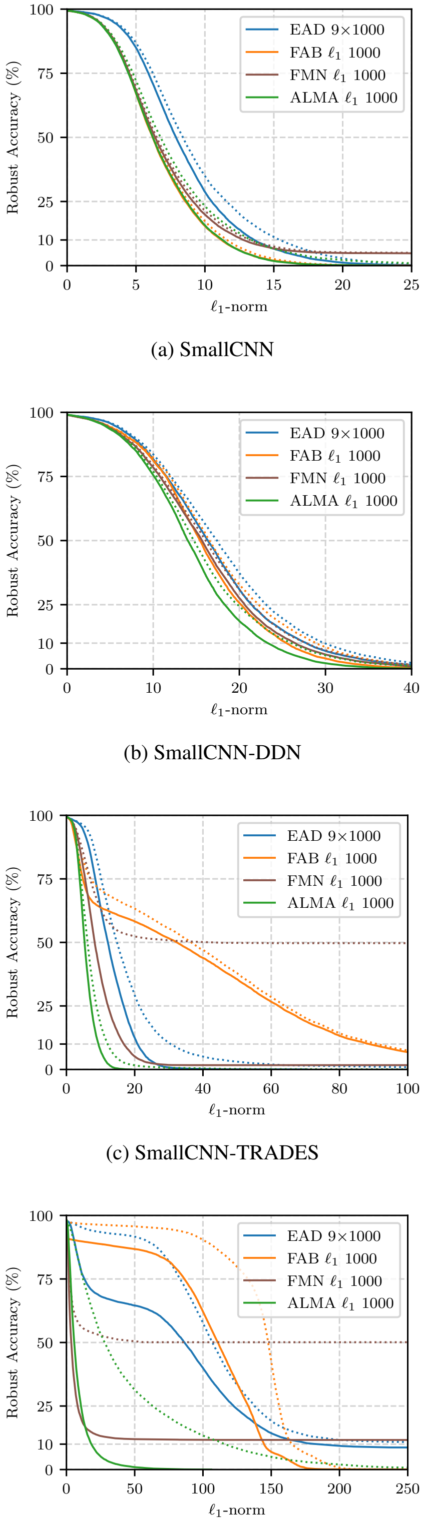
Table 13: Performance of the CIEDE2000 attacks on the CIFAR10 dataset for each model.

Model	Attack	ASR (%)	Median LPIPS $\times 10^{-2}$	forwards / backwards
RN-50	C&W LPIPS 9×1000	100	2.39	2 332 / 2 325
	LPA‡ [21]	100	5.02	1 159 / 1 149
	ALMA LPIPS 100	100	0.34	100 / 100
	ALMA LPIPS 1000	100	0.24	1 000 / 1 000
RN-50 ℓ_2-AT	C&W LPIPS 9×1000	100	2.22	7 701 / 7 700
	LPA‡ [21]	100	6.83	1 257 / 1 247
	ALMA LPIPS 100	100	3.96	100 / 100
	ALMA LPIPS 1000	100	2.90	1 000 / 1 000
RN-50 ℓ_∞-AT	C&W LPIPS 9×1000	100	1.68	6 753 / 6 750
	LPA‡ [21]	100	5.66	1 218 / 1 208
	ALMA LPIPS 100	100	2.99	100 / 100
	ALMA LPIPS 1000	100	2.11	1 000 / 1 000

Table 14: Performance of the LPIPS variant of ALMA on the ImageNet dataset for each model. ‡A binary search is performed on each sample to get a minimal perturbation attack (Equation 2).

Model	Attack	ASR (%)	Median $1-$SSIM $\times 10^{-5}$	forwards / backwards
RN-50	ALMA SSIM 100	100	0.44	100 / 100
	ALMA SSIM 1000	100	0.05	1 000 / 1 000
RN-50 ℓ_2-AT	ALMA SSIM 100	100	16.13	100 / 100
	ALMA SSIM 1000	100	5.58	1 000 / 1 000
RN-50 ℓ_∞-AT	ALMA SSIM 100	100	8.69	100 / 100
	ALMA SSIM 1000	100	2.77	1 000 / 1 000

Table 15: Performance of the SSIM variant of ALMA on the ImageNet dataset for each model.



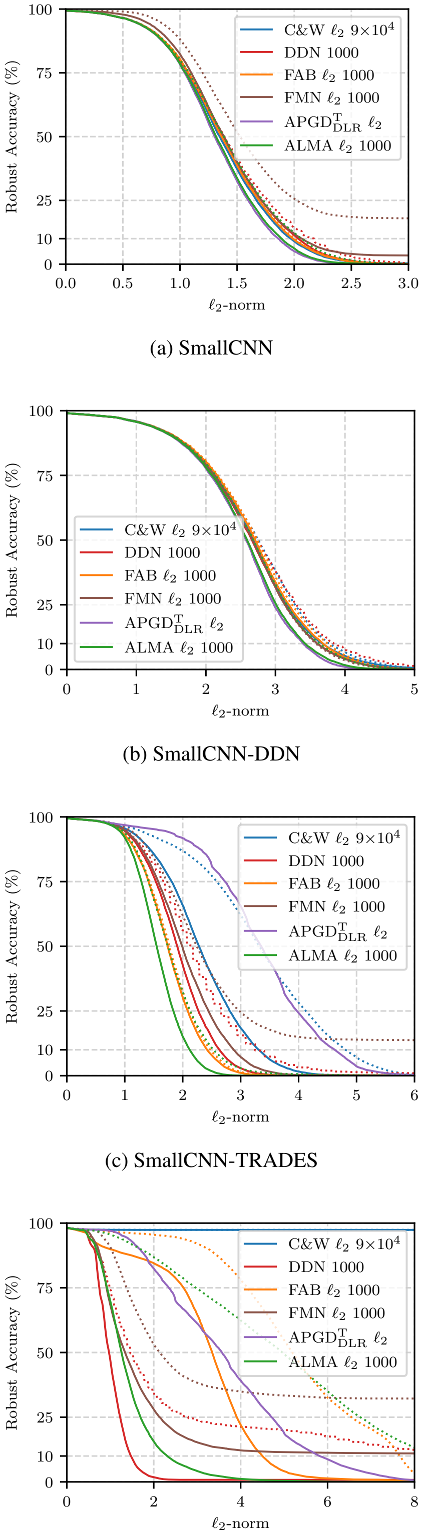
(a) SmallCNN

(b) SmallCNN-DDN

(c) SmallCNN-TRADES

(d) CROWN-IBP

Figure 6: Robust accuracy curves for MNIST models against ℓ_1 attacks.



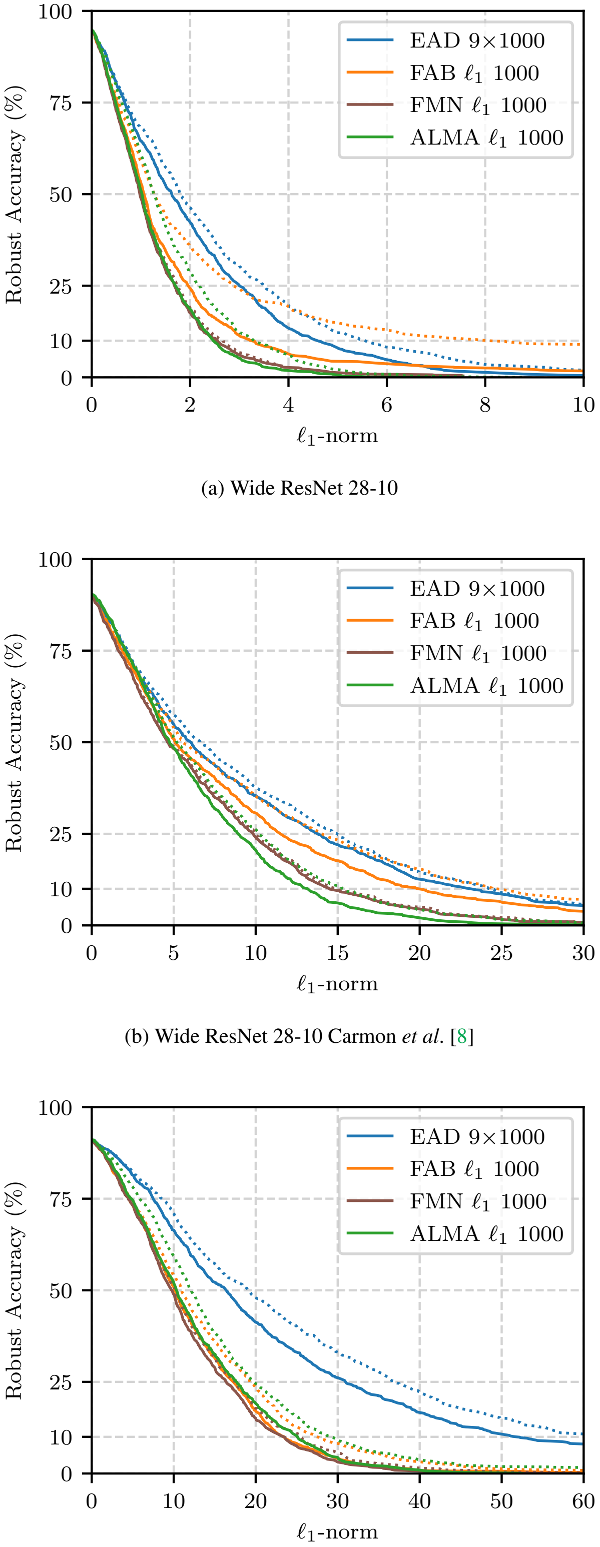
(a) SmallCNN

(b) SmallCNN-DDN

(c) SmallCNN-TRADES

(d) CROWN-IBP

Figure 7: Robust accuracy curves for MNIST models against ℓ_2 attacks.

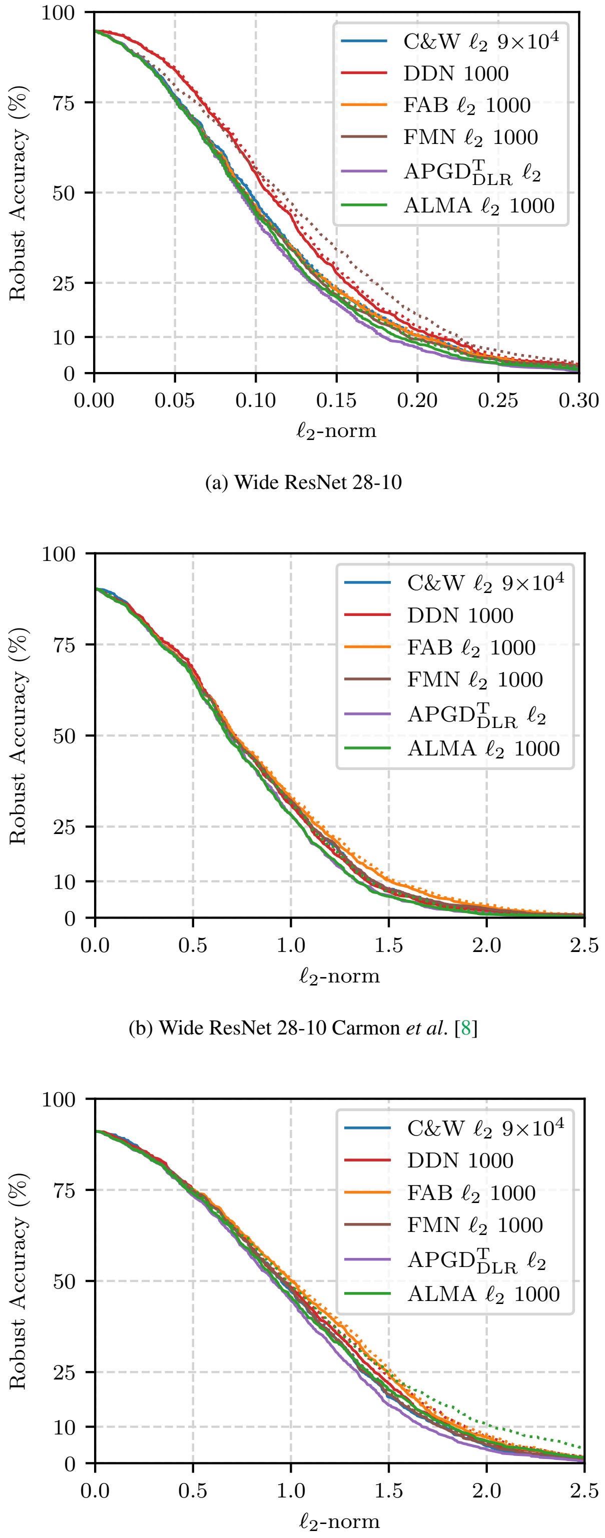


(a) Wide ResNet 28-10

(b) Wide ResNet 28-10 Carmon *et al.* [8]

(c) ResNet-50 Augustin *et al.* [1]

Figure 8: Robust accuracy curves for CIFAR10 models against ℓ_1 attacks.

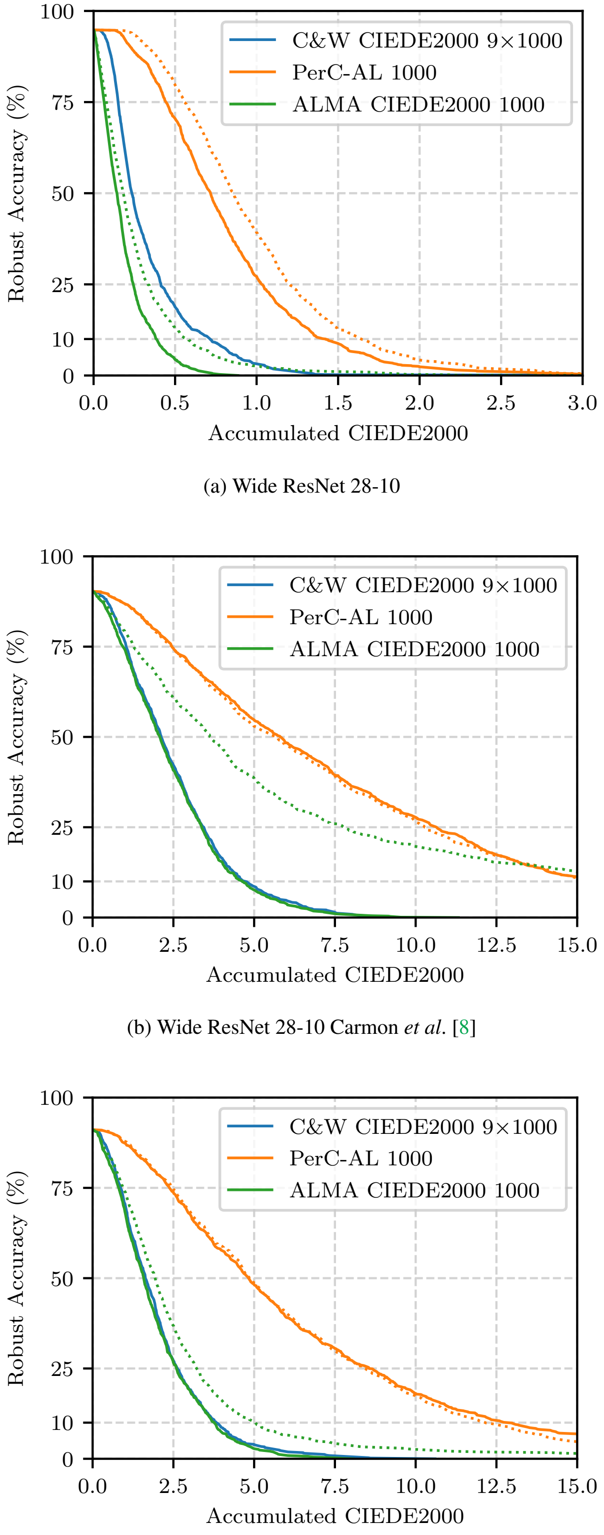


(a) Wide ResNet 28-10

(b) Wide ResNet 28-10 Carmon *et al.* [8]

(c) ResNet-50 Augustin *et al.* [1]

Figure 9: Robust accuracy curves for CIFAR10 models against ℓ_2 attacks.

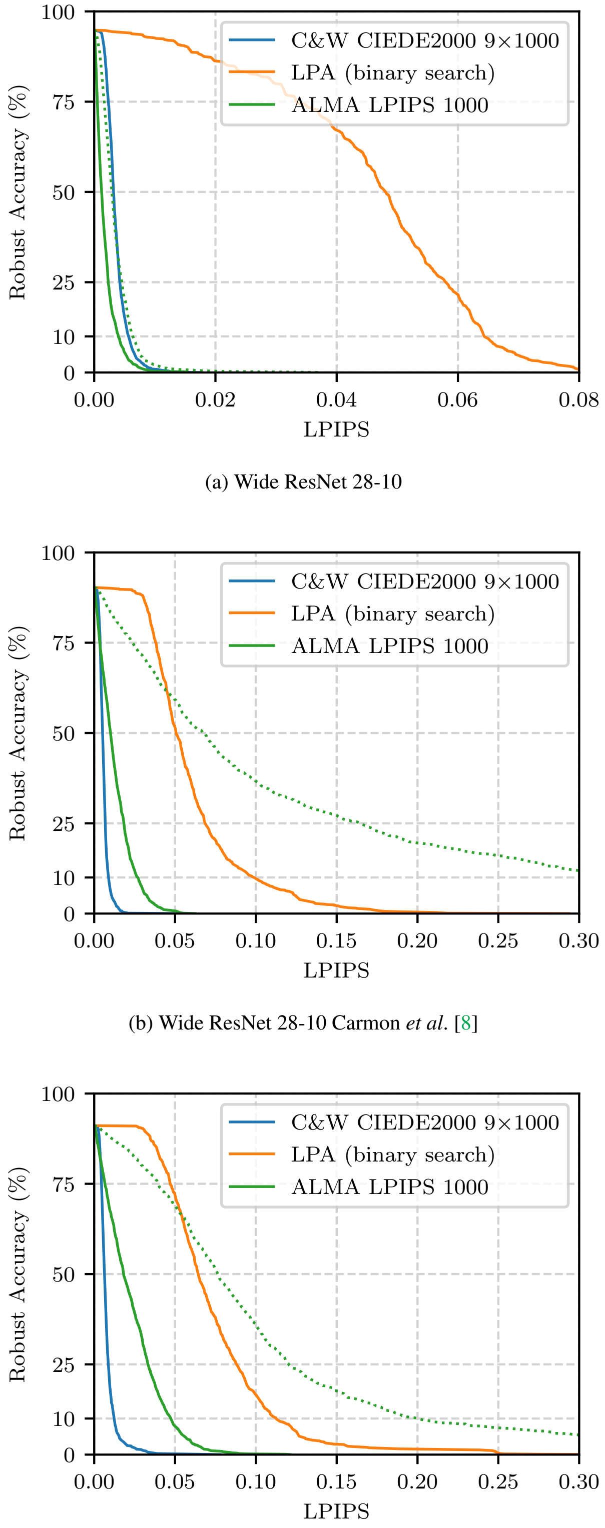


(a) Wide ResNet 28-10

(b) Wide ResNet 28-10 Carmon *et al.* [8]

(c) ResNet-50 Augustin *et al.* [1]

Figure 10: Robust accuracy curves for CIFAR10 models against CIEDE2000 attacks.

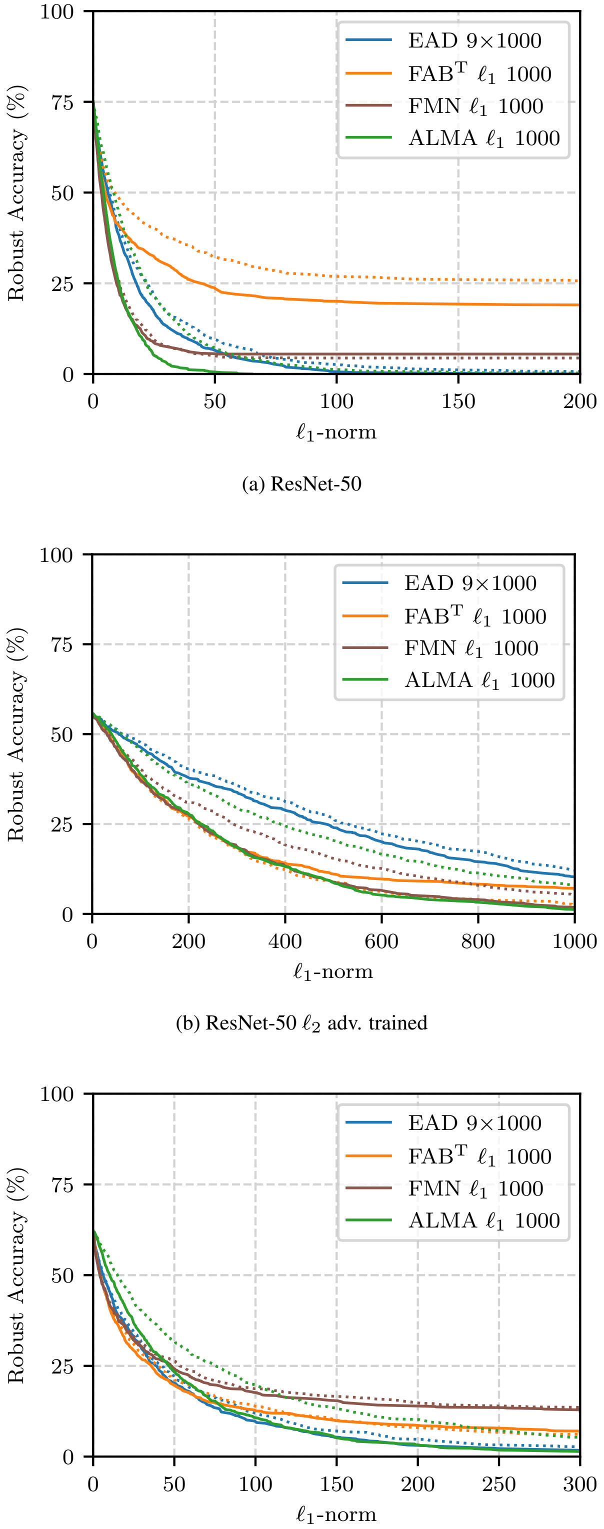


(a) Wide ResNet 28-10

(b) Wide ResNet 28-10 Carmon *et al*. [8]

(c) ResNet-50 Augustin *et al*. [1]

Figure 11: Robust accuracy curves for CIFAR10 models against LPIPS attacks.

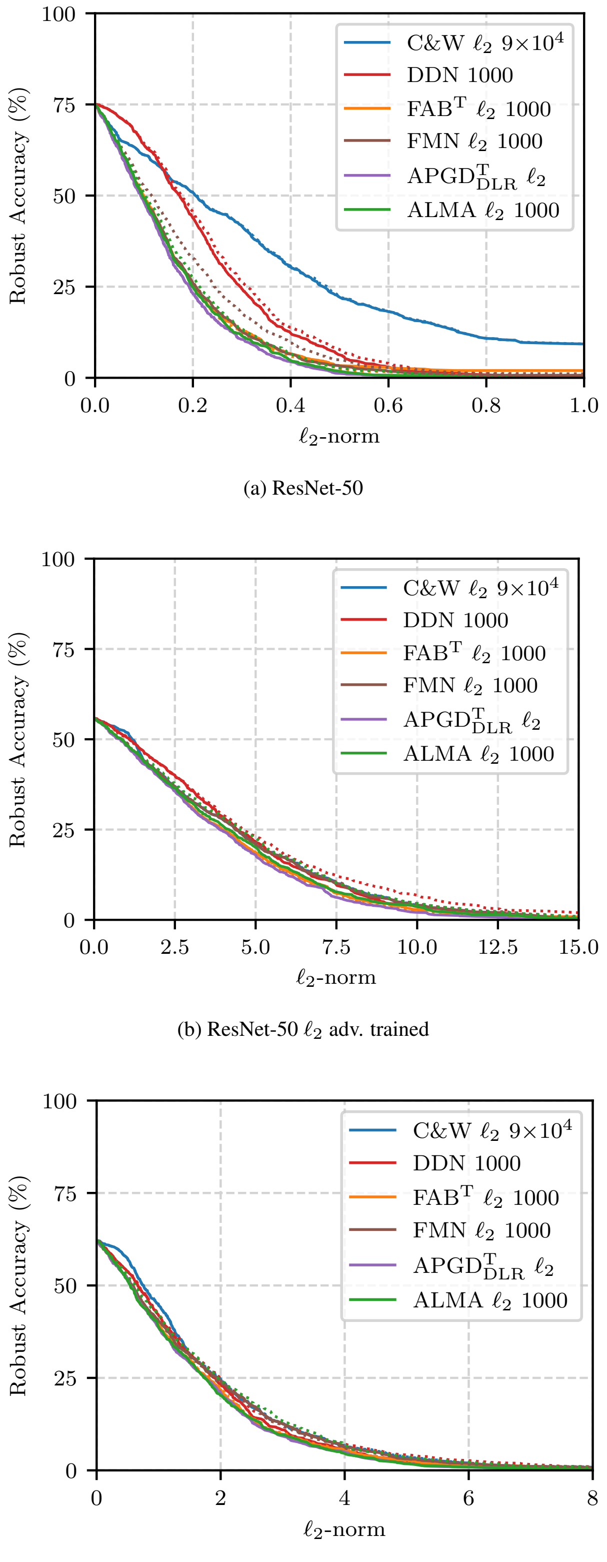


(a) ResNet-50

(b) ResNet-50 ℓ_2 adv. trained

(c) ResNet-50 ℓ_∞ adv. trained

Figure 12: Robust accuracy curves for ImageNet models against ℓ_1 attacks.

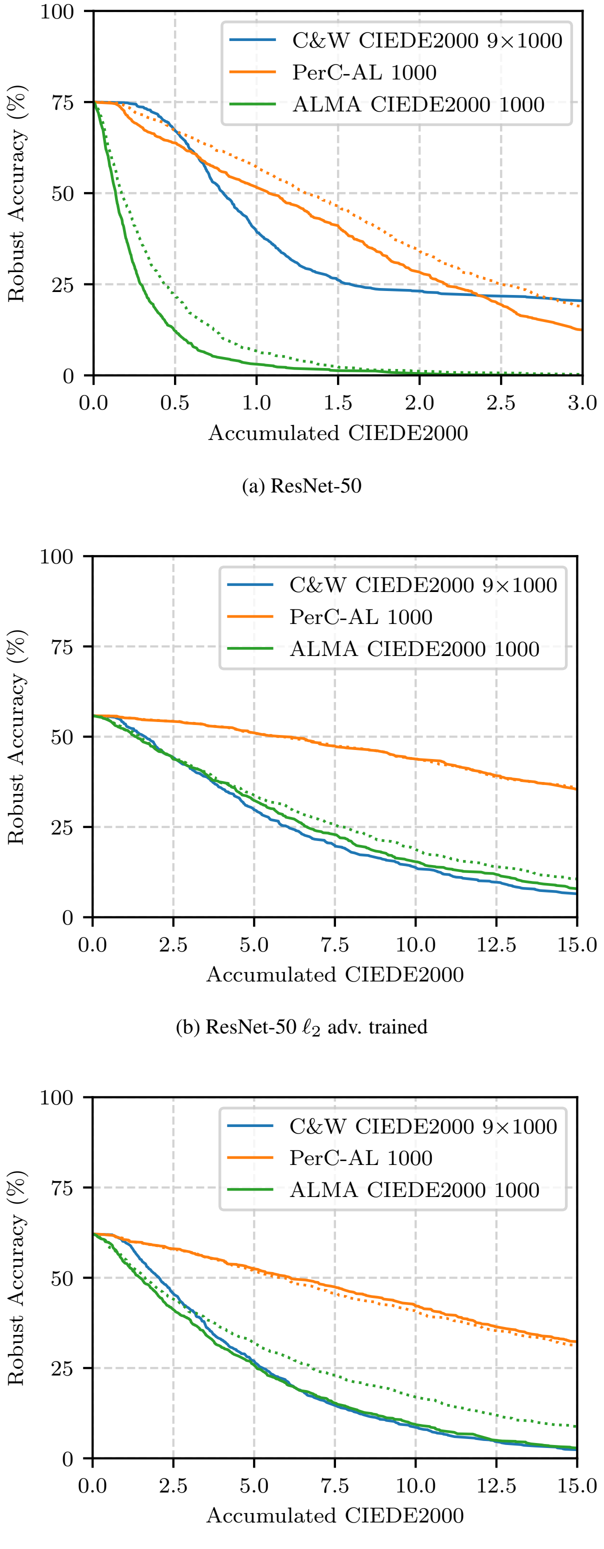


(a) ResNet-50

(b) ResNet-50 ℓ_2 adv. trained

(c) ResNet-50 ℓ_∞ adv. trained

Figure 13: Robust accuracy curves for ImageNet models against ℓ_2 attacks.

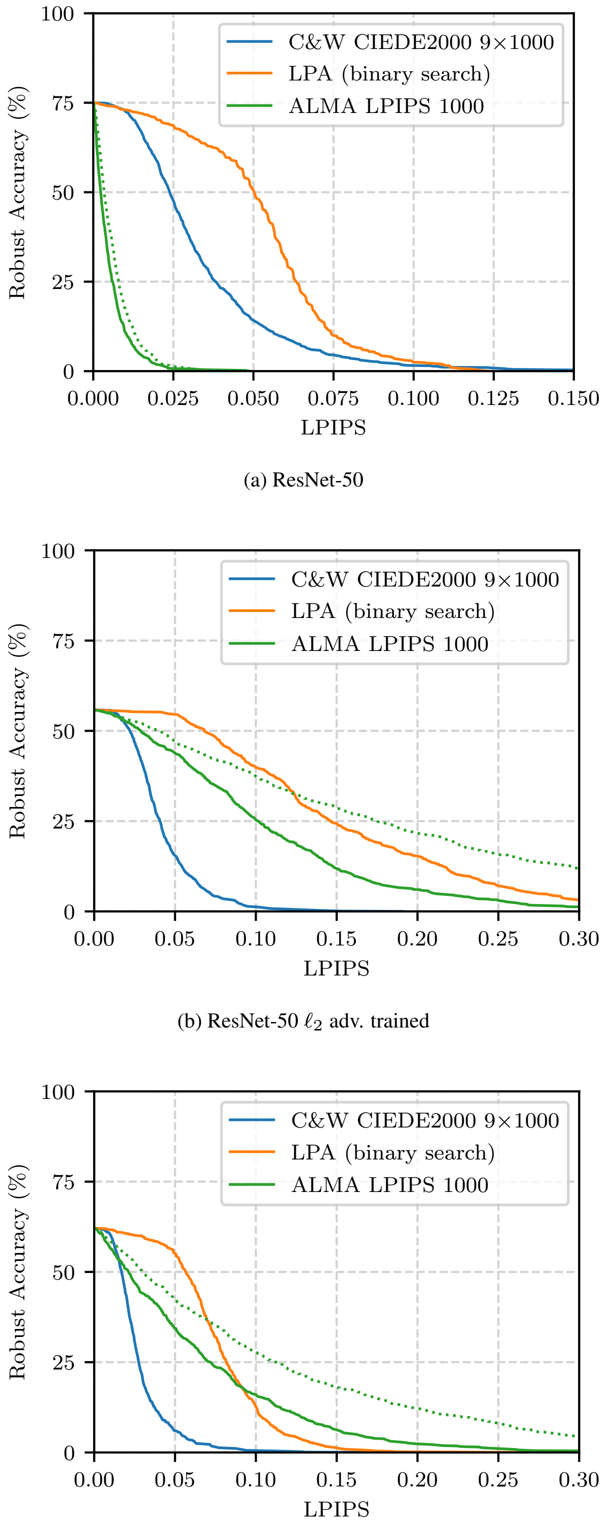


(a) ResNet-50

(b) ResNet-50 ℓ_2 adv. trained

(c) ResNet-50 ℓ_∞ adv. trained

Figure 14: Robust accuracy curves for ImageNet models against CIEDE2000 attacks.



(a) ResNet-50

(b) ResNet-50 ℓ_2 adv. trained

(c) ResNet-50 ℓ_∞ adv. trained

Figure 15: Robust accuracy curves for ImageNet models against LPIPS attacks.