

Table 1: Table presents accuracy of non-targeted attacks on DNN classifiers after being attacked by each method along with the l_2 distance and non-normalized residual of the adversarial patch. We provide the mean and variance of each metric over 10 runs.

Attack Method	VGG-16			ResNet-50		
	Accuracy	l_2	Non-Normalized Residual	Accuracy	l_2	Non-Normalized Residual
-	73.36%	-	-	76.12%	-	-
CamoPatch	9.70% (0.03)	0.09 (0.02)[†]	0.11 (0.02)[†]	10.00% (0.02)[†]	0.08 (0.01)[†]	0.10 (0.01)[†]
Patch-RS*	6.82% (0.04)	0.42 (0.02) [‡]	0.30 (0.05) [‡]	15.92% (0.02) [‡]	0.45 (0.04) [‡]	0.31 (0.04) [‡]
Patch-RS	6.82% (0.04)	0.63 (0.01) [‡]	0.61 (0.07) [‡]	15.92% (0.02) [‡]	0.67 (0.08) [‡]	0.69 (0.07) [‡]
TPA	47.11% (1.30) [‡]	0.61 (0.13) [‡]	0.55 (0.05) [‡]	38.98% (1.41) [‡]	0.61 (0.07) [‡]	0.58(0.07) [‡]
OPA	32.19% (0.10) [‡]	0.71 (0.20) [‡]	0.64 (0.06) [‡]	27.91% (1.12) [‡]	0.71 (0.14) [‡]	0.66 (0.04) [‡]
LOAP	37.99% (0.40) [‡]	0.68 (0.02) [‡]	0.63 (0.05) [‡]	47.99% (0.10) [‡]	0.78 (0.12) [‡]	0.67 (0.05) [‡]
Adv-watermark	32.00% (0.10) [‡]	0.13(0.08) [‡]	0.25(0.05) [‡]	35.00% (0.40) [‡]	0.16(0.01) [‡]	0.31(0.07) [‡]

Attack Method	AT-WideResNet-50-2)			AT-ResNet-50		
	Accuracy	l_2	Non-Normalized Residual	Accuracy	l_2	Non-Normalized Residual
-	68.46%	-	-	64.02%	-	-
CamoPatch	12.98% (0.01)[†]	0.14 (0.05)[†]	0.12 (0.07)[†]	6.00% (0.03)[†]	0.15 (0.03)[†]	0.13 (0.03)[†]
Patch-RS*	14.42% (0.01) [‡]	0.43 (0.07) [‡]	0.30 (0.05) [‡]	12.00% (0.02) [‡]	0.41 (0.12) [‡]	0.33 (0.05) [‡]
Patch-RS	14.42% (0.01) [‡]	0.74 (0.08) [‡]	0.42 (0.07) [‡]	12.00% (0.02) [‡]	0.74 (0.09) [‡]	0.43 (0.07) [‡]
TPA	51.66% (1.3) [‡]	0.82 (1.21) [‡]	0.82 (0.07) [‡]	34.82% (1.41) [‡]	0.92 (0.05) [‡]	0.87(0.09) [‡]
OPA	36.88% (0.1) [‡]	0.76 (0.20) [‡]	0.74 (0.05) [‡]	24.83% (1.12) [‡]	0.77 (0.14) [‡]	0.75 (0.04) [‡]
LOAP	38.85% (0.4) [‡]	0.56 (0.02) [‡]	0.46 (0.03) [‡]	48.89% (0.1) [‡]	0.72 (0.18) [‡]	0.64 (0.03) [‡]
Adv-watermark	52.00% (0.3) [‡]	0.37(0.05) [‡]	0.23(0.07) [‡]	44.00% (0.3) [‡]	0.42 (0.02) [‡]	0.29 (0.07) [‡]

Attack Method	ViT-B/16			BagNet9 with PatchGuard		
	Accuracy	l_2	Non-Normalized Residual	Accuracy	l_2	Non-Normalized Residual
-	77.91%	-	-	55.1%	-	-
CamoPatch	8.00% (0.05)[†]	0.09 (0.02)	0.12 (0.02)	3.20% (0.01)[†]	0.07(0.03)[†]	0.11 (0.01)[†]
Patch-RS*	19.00% (0.10) [‡]	0.68 (0.05) [‡]	0.39 (0.07) [‡]	5.80% (0.02) [‡]	0.42 (0.05) [‡]	0.30 (0.05) [‡]
Patch-RS	19.00% (0.10) [‡]	0.71 (0.12) [‡]	0.41 (0.09) [‡]	5.80% (0.02) [‡]	0.62 (0.18) [‡]	0.57 (0.11) [‡]
TPA	38.12% (0.91) [‡]	0.59 (0.08) [‡]	0.54 (0.09) [‡]	32.87% (1.45) [‡]	0.62 (0.11) [‡]	0.61(0.09) [‡]
OPA	33.09% (0.17) [‡]	0.68 (0.23) [‡]	0.68 (0.07) [‡]	57.89% (2.01) [‡]	0.61 (0.16) [‡]	0.67 (0.04) [‡]
LOAP	43.91% (0.80) [‡]	0.63 (0.05) [‡]	0.50 (0.13) [‡]	72.82% (0.14) [‡]	0.89 (0.23) [‡]	0.78 (0.11) [‡]
Adv-watermark	36.01% (0.12) [‡]	0.17(0.04) [‡]	0.28(0.03) [‡]	42.00% (0.45) [‡]	0.14(0.01) [‡]	0.29(0.05) [‡]

[†] denotes the performance of the method significantly outperforms the compared methods according to the Wilcoxon signed-rank test at the 5% significance level; [‡] denotes the corresponding method is significantly outperformed by the best performing method (shaded).

Table 2: Table presents accuracy of targeted attacks on DNN classifiers after being attacked by each method along with the l_2 distance and non-normalized residual of the adversarial patch. We provide the mean and variance of each metric over 10 runs.

Attack Method	VGG-16			AT-ResNet-50		
	Accuracy	l_2	Non-Normalized Residual	Accuracy	l_2	Non-Normalized Residual
-	73.36%	-	-	68.46%	-	-
CamoPatch	32.82% (3.20)[†]	0.22 (0.12)[†]	0.2 (0.03)[†]	58.28% (4.03)[†]	0.34 (0.02)[†]	0.28 (0.05)[†]
Patch-RS*	41.49% (5.10) [‡]	0.47 (0.27) [‡]	0.32 (0.06) [‡]	90% (2.4) [‡]	0.73 (0.01) [‡]	0.42 (0.07) [‡]
Patch-RS	41.49% (5.10) [‡]	0.75 (0.04) [‡]	0.47 (0.02) [‡]	90% (2.4) [‡]	0.78 (0.01) [‡]	0.43 (0.02) [‡]

[†] denotes the performance of the method significantly outperforms the compared methods according to the Wilcoxon signed-rank test at the 5% significance level; [‡] denotes the corresponding method is significantly outperformed by the best performing method (shaded).

l_i	t	N	σ	VGG-16			
				ASR	l_2	SSIM	Runtime(s)
-	-	-	-	76.12%	-	-	-
1	300	100	0.1	87.12%(1.0)	0.09(0.04)	0.98(0.00)	440.03(10.32)
4	100	100	0.1	89.21%(1.5)	0.09(0.07)	0.98(0.00)	440.01(10.05)
4	300	100	0.1	90.36%(1.0)	0.09(0.05)	0.98(0.00)	440.03(10.32)
4	300	100	0.3	88.34%(2.0)	0.09(0.06)	0.98(0.00)	439.13(10.56)

Table 3: Table presents the attack statistics of the top four performing parameters configurations (in terms of attack success rate (ASR)) of the proposed method. The mean and variance of each metric is calculated over 5 runs. According to the Wilcoxon signed-rank test, no configuration is able to outperform another.

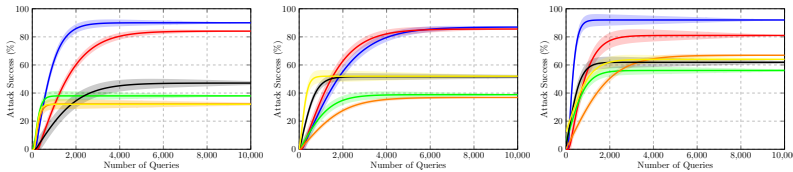


Figure 1: Success Attack success over queries for non-targeted attacks. Figure shows the convergence plots of CamoPatch (blue), Patch-RS* (Red), Adv-watermark (yellow), TPA (black), OPA (green) and LOAP (orange).