ATAC: Augmentation-Based Test-Time Adversarial Correction for CLIP

Supplementary Material

10. Results Against Other Attacks

To further validate the generality and reliability of ATAC, we extend our evaluation beyond the standard PGD setting ($\epsilon = 4/255$) to two widely recognized and complementary benchmarks: AutoAttack [10] and the Carlini–Wagner (CW) attack [5]. We use the "plus" version of AutoAttack that integrates six attacks, including both targeted and untargeted, as well as gradient-based and gradient-free attacks, in order to provide a standardized and rigorous robustness evaluation. In contrast, the CW attack formulates adversarial example generation as an explicit optimization problem that seeks minimal perturbations leading to confident misclassification, making it a representative test of fine-grained vulnerability beyond gradient-based methods. We evaluate both AutoAttack and CW under two perturbation budgets, $\epsilon \in 1/255, 4/255$, to examine robustness under both mild and strong attack regimes. We further evaluate PGD with a budget of $\epsilon = 1/255$.

As shown in Tables 5 and 6, ATAC consistently achieves large gains in robust accuracy across all datasets and attack settings, while maintaining nearly unchanged clean performance. Even under strong attacks such as CW or AutoAttack at higher ϵ, ATAC restores model predictions to a level comparable to or exceeding the clean baseline, highlighting its ability to generalize beyond PGD and effectively counter diverse adversaries.

Overall, these results confirm that **ATAC is not attack-specific**: it maintains strong and consistent robustness under a wide range of threat models, demonstrating its potential as a general-purpose test-time defense mechanism.

11. On the Distribution of Consistency-Scores

In Sec. 4.2 we argue that the augmentation-induced latent drift vectors are scattered for clean samples and consistent for adversarial inputs. To verify our claim, we analyze the distribution of τ-scores for clean and adversarial inputs, and report the separability of the two distributions.

The last column of Fig. 3 shows the separability of clean and adversarial τ-distributions using our set of augmentations. Our setting achieves a consistently high area under the curve (AUC) of nearly 1 in all cases, demonstrating that adversarial and clean inputs can be effectively separated using the consistency of their augmentation-induced latent drifts.

12. Further Ablations

In Sec. 5.3, we find that the effect of α is minimal while τ^* is crucial. In this section, we investigate the effect of different augmentation choices. To understand which aspects of augmentations contribute to performance, we construct five ablation settings.

- *default*: the original setting used in our main experiments.
- *asymmetric*: when initially selecting augmentations, we hypothesized that averaging drift vectors of symmetric augmentations could reduce introduced bias. This setting is used to validate that hypothesis. The augmentations in this setting are horizontal flip, and rotations with degrees +15, -20, -25, +30.
- *random*: we replace the deterministic augmentations (horizontal flip with $p = 1$ and fixed-degree rotations) in the default setting with random flips and rotations with a probability of $p = 0.5$.
- *color*: replaces flip and rotation with color jittering augmentations. We used five random color jittering transformations, with brightness ± 40, contrast ± 40, saturation ± 40, and hue ± 15.
- *more*: we include both horizontal and vertical flips, as well as 8 different rotations with degrees ± 15, ± 20, ± 25, and ± 30.

As shown in Fig. 7, there is only a negligible difference between *default* and *asymmetric*, indicating that symmetry does not necessarily improve performance. Moreover, varying rotation degrees can even yield improvements, offering more flexibility for augmentation choices in deployment. The *random* setting does not achieve robustness comparable to the first two settings, although a moderate gain still exists. We hypothesize this is due to insufficient augmentation; extending the range could mitigate this deficiency but would also introduce instability and potentially degrade performance. The *color* setting yields the poorest performance, which is consistent with the finding in [11] that CLIP's representations are most affected by noise addition, followed by color-variant transformations (including color jitter). This also suggests that ATAC relies on label-preserving augmentations, while those that introduce substantial embedding shifts (e.g. noise addition, blur, coarse dropout...[11]) may be less suitable. Finally, although the *more* setting attains the highest clean accuracy, it yields roughly 10% lower robust accuracy compared to *default*, indicating that simply adding more augmentations does not necessarily lead to consistent gains. In practice, this shows that ATAC does not need many costly augmentations in deployment, as a small number of transformations already delivers high performance.

Dataset		No Defense				ATAC			
		$auto_1$	$auto_4$	CW_1	CW_4	$auto_1$	$auto_4$	CW_1	CW_4
CIFAR10	Rob.	0.01	0.01	0.79	0.00	84.72 (+84.71)	85.18 (+85.17)	79.24 (+78.45)	**91.58** (+91.58)
	Acc.	**85.08**	85.08	85.08	85.08	81.04 (-4.04)	81.04 (-4.04)	81.04 (-4.04)	81.04 (-4.04)
CIFAR100	Rob.	0.11	0.11	0.30	0.00	56.77 (+56.66)	57.49 (+57.38)	53.75 (+53.45)	**78.08** (+78.08)
	Acc.	**57.20**	57.20	57.20	57.20	53.74 (-3.46)	53.74 (-3.46)	53.74 (-3.46)	53.74 (-3.46)
STL10	Rob.	0.00	0.00	11.86	0.01	96.26 (+96.26)	96.39 (+96.39)	90.42 (+78.56)	**98.01** (+98.00)
	Acc.	**96.42**	96.42	96.42	96.42	95.72 (-0.70)	95.72 (-0.70)	95.72 (-0.70)	95.72 (-0.70)
Flowers102	Rob.	0.02	0.02	1.51	0.00	64.12 (+64.10)	64.79 (+64.77)	48.77 (+47.26)	**84.92** (+84.92)
	Acc.	**65.56**	65.56	65.56	65.56	65.34 (-0.22)	65.34 (-0.22)	65.34 (-0.22)	65.34 (-0.22)
FGVCAircraft	Rob.	0.09	0.09	0.00	0.00	15.06 (+14.97)	17.52 (+17.43)	19.53 (+19.53)	**54.10** (+54.10)
	Acc.	**20.16**	20.16	20.16	20.16	19.83 (-0.33)	19.83 (-0.33)	19.83 (-0.33)	19.83 (-0.33)
DTD	Rob.	0.16	0.16	2.55	0.05	38.94 (+38.78)	40.00 (+39.84)	39.79 (+37.24)	**64.73** (+64.68)
	Acc.	**40.11**	40.11	40.11	40.11	39.15 (-0.96)	39.15 (-0.96)	39.15 (-0.96)	39.15 (-0.96)
Avg.	Rob.	0.07	0.07	2.84	0.01	59.31 (+59.24)	60.23 (+60.16)	55.25 (+52.41)	**78.57** (+78.56)
	Acc.	**60.76**	60.76	60.76	60.76	59.14 (-1.62)	59.14 (-1.62)	59.14 (-1.62)	59.14 (-1.62)

Table 5. ATAC under various attacks. Here, *auto* denotes AutoAttack and *CW* denotes the Carlini–Wagner attack. The subscript indicates the attack budget ϵ, e.g., $auto_1$ corresponds to AutoAttack with $\epsilon = 1/255$. For AutoAttack, we adopt the "plus" version, which integrates untargeted attacks (APGD-CE, APGD-DLR, FAB), targeted attacks (APGD-T, FAB-T), and a gradient-free attack (Square), thereby providing a comprehensive and reliable evaluation of adversarial robustness.

13. Adaptive Attack Algorithms

Here, we give the full pseudocodes for our attacks. The adaptive attack against our method is given in Algorithm 1, and the adaptive attack against TTC is given in Algorithm 2. In both pseudocodes, we use $\text{pred}(\cdot, \cdot)$ as a shorthand for the calculation of class-wise logits (see Eq. (1)). σ denotes the sigmoid function. As in the main text, we omit denoting the projection of adversarial attacks to the input space for the sake of simplicity.

For the adaptive attack against ATAC, we used $\epsilon = 4/255$, $\gamma = 1/255$, $\mathcal{K} = 40$, $\lambda = 1$, and $T = 10$ optimization steps.

Similarly, for the adaptive attack against TTC, we used $\epsilon = 4/255$, $\gamma = 1/255$, $\mathcal{K} = 40$, $\lambda = 1$, and $T = 10$ optimization steps. The parameters of TTC were $\epsilon_{ttc} = 2/255$, $\eta = 1/255$, $\epsilon_\tau = 2/255$, and $\tau_{thresh} = 0.2$.

In both cases, due to the large value of the gating temperature $\mathcal{K}$, the soft correcton and soft counterattack parts of our attacks can be interpreted as "nearly hard", and the hard variant would yield highly similar results.

(%)		CLIP	Adversarial Finetuning				Test-time Defense				Δ
			CLIP-FT	TeCoA	PMG-AFT	FARE	TTE	TTC	R-TPT	ATAC (ours)	
CIFAR10	Rob.	0.74	3.34	33.61	40.66	19.65	41.35 ± 6.14	28.75 ± 0.18	70.80	**81.03**	+66.36
	Acc.	85.12	**84.90**	64.61	70.69	74.44	84.74 ± 0.40	81.18 ± 0.07	82.19	81.03	−4.09
CIFAR100	Rob.	0.26	0.90	18.95	22.52	11.40	20.06 ± 4.03	14.31 ± 0.25	43.85	**64.24**	+63.98
	Acc.	57.14	**59.51**	35.96	40.32	46.67	58.61 ± 0.25	56.34 ± 0.20	52.69	53.64	−3.50
STL10	Rob.	11.0	12.73	70.08	73.08	59.06	78.48 ± 3.83	76.70 ± 0.23	**90.59**	90.41	+79.41
	Acc.	96.40	94.49	87.40	88.56	91.72	**96.26 ± 0.04**	95.85 ± 0.04	96.09	95.71	−0.69
Caltech101	Rob.	14.67	14.21	55.51	61.08	50.74	67.56 ± 3.88	65.78 ± 0.07	**79.32**	72.41	+57.74
	Acc.	85.66	83.63	71.68	75.45	80.95	85.84 ± 0.09	86.53 ± 0.07	**86.62**	85.14	−0.52
Caltech256	Rob.	8.47	6.76	43.19	45.91	38.79	60.09 ± 4.03	60.11 ± 0.04	67.51	**68.02**	+59.55
	Acc.	81.72	78.53	61.14	62.24	73.32	**82.49 ± 0.08**	79.66 ± 0.04	77.67	80.72	−1.00
OxfordPets	Rob.	1.04	2.10	38.35	41.18	31.07	50.33 ± 7.30	57.87 ± 0.15	71.79	**77.11**	+76.07
	Acc.	87.44	84.14	62.12	65.88	79.37	**88.13 ± 0.13**	83.35 ± 0.21	84.46	87.30	−0.14
Flowers102	Rob.	1.14	0.54	21.94	23.43	17.14	35.88 ± 4.72	39.14 ± 0.28	52.07	**54.74**	+53.60
	Acc.	65.46	53.37	36.80	37.00	47.98	65.18 ± 0.22	64.16 ± 0.19	62.92	**65.34**	−0.12
FGVCAircraft	Rob.	0.00	0.00	2.49	2.22	1.35	6.23 ± 1.37	13.77 ± 0.38	13.62	**23.37**	+23.37
	Acc.	20.10	14.04	5.31	5.55	10.86	**20.19 ± 0.36**	18.00 ± 0.16	19.14	19.80	−0.30
StanfordCars	Rob.	0.02	0.06	8.76	11.65	6.75	22.36 ± 4.17	33.01 ± 0.07	**43.75**	27.12	+27.10
	Acc.	52.02	42.11	20.91	25.44	38.68	**52.73 ± 0.31**	48.16 ± 0.16	61.75	51.45	−0.57
Country211	Rob.	0.04	0.03	1.78	2.12	0.85	3.05 ± 0.89	7.09 ± 0.04	8.80	**30.14**	+30.10
	Acc.	15.25	12.07	4.75	4.64	9.26	14.66 ± 0.16	13.08 ± 0.05	13.40	**16.52**	+1.27
Food101	Rob.	0.70	0.42	13.90	18.57	11.65	43.94 ± 6.97	57.84 ± 0.15	68.04	**76.47**	+75.77
	Acc.	83.88	64.86	29.98	36.61	55.31	**83.96 ± 0.02**	82.18 ± 0.02	83.41	83.57	−0.31
EuroSAT	Rob.	0.03	0.04	11.96	12.60	10.67	6.91 ± 2.13	12.19 ± 0.24	14.16	**66.90**	+66.87
	Acc.	42.59	27.64	16.58	18.53	21.88	44.38 ± 1.60	53.24 ± 0.09	21.83	38.32	−4.21
DTD	Rob.	2.98	2.39	17.61	14.95	15.64	23.90 ± 2.34	27.32 ± 0.25	34.10	**47.93**	+44.95
	Acc.	40.64	36.49	25.16	21.76	32.07	**41.33 ± 0.32**	36.98 ± 0.21	42.66	39.15	−1.49
Avg.	Rob.	3.16	3.35	26.01	28.46	21.14	35.40	37.99	50.65	**59.99**	+56.83
	Acc.	62.57	56.60	40.18	42.51	50.96	**62.96**	61.44	60.37	61.37	−1.20

Table 6. Classification accuracy (%) on both adversarial images (Rob.) under 10-step PGD attack at $\epsilon_a = 1/255$ and clean images (Acc.) across datasets. Finetuning-based models are implemented as references. For test-time methods, we report mean±std over 3 runs.

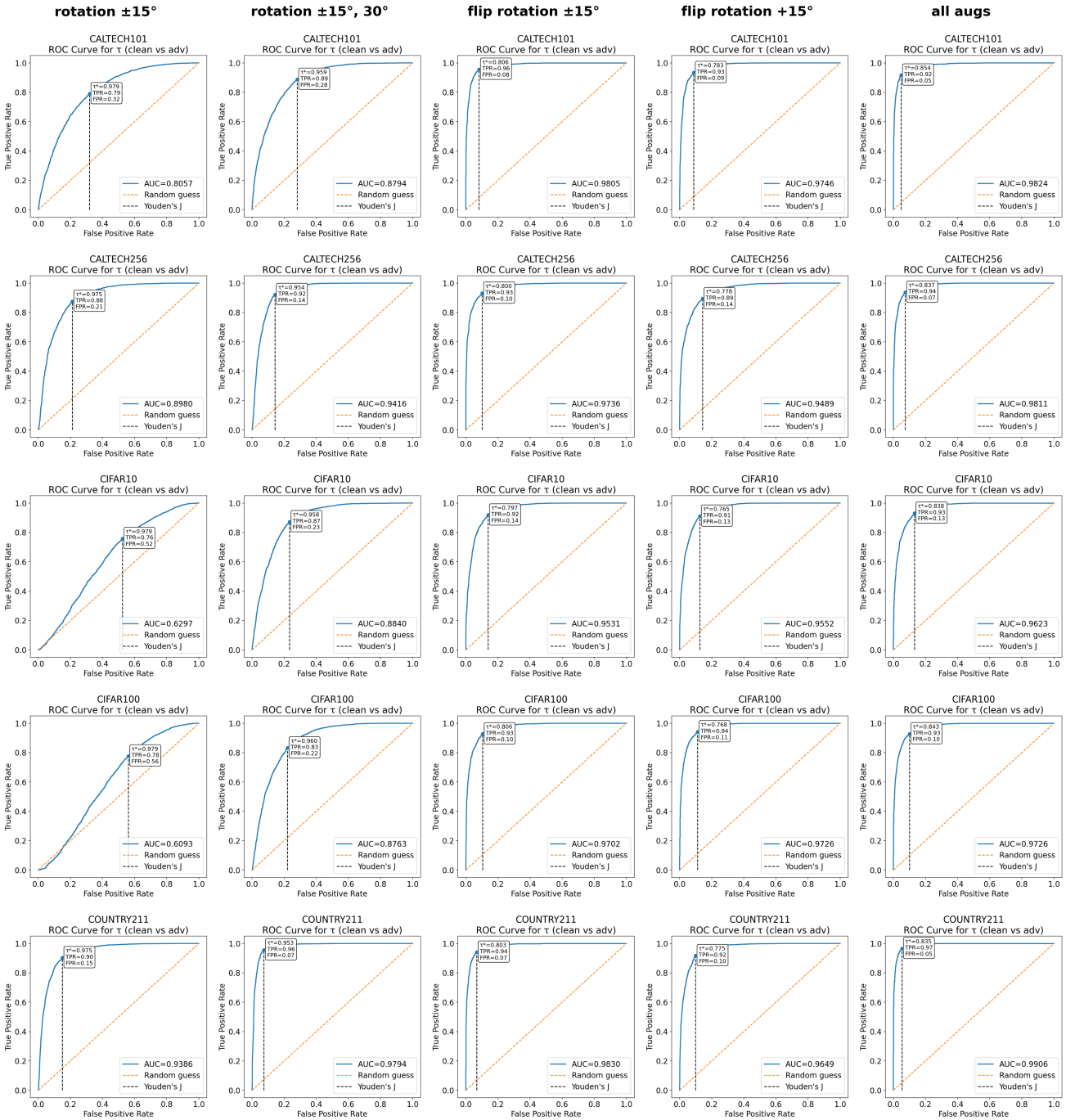

Figure 3. ROC curves of τ-scores of different augmentation settings on different datasets.

Dataset		No Defense	ATAC				
			default	*asymmetric*	*random*	*color*	*more*
STL10	Rob.	0.00	97.94 (+97.94)	**98.00** (+98.00)	41.06 (+41.06)	2.69 (+2.69)	81.94 (+81.94)
	Acc.	96.19	95.38 (-0.81)	95.44 (-0.75)	95.81 (-0.38)	95.56 (-0.63)	**96.19** (+0.00)
Caltech101	Rob.	0.00	67.63 (+67.63)	**67.81** (+67.81)	31.13 (+31.13)	6.25 (+6.25)	55.94 (+55.94)
	Acc.	68.38	67.69 (-0.69)	67.38 (-1.00)	68.31 (-0.07)	**68.38** (+0.00)	**68.38** (+0.00)
OxfordPets	Rob.	0.00	95.56 (+95.56)	**95.88** (+95.88)	42.25 (+42.25)	2.44 (+2.44)	84.88 (+84.88)
	Acc.	83.13	83.25 (+0.12)	83.13 (+0.00)	**83.31** (+0.18)	82.81 (-0.32)	83.19 (+0.06)
Flowers102	Rob.	0.00	**84.69** (+84.69)	84.19 (+84.19)	39.06 (+39.06)	2.13 (+2.13)	81.25 (+81.25)
	Acc.	65.19	64.81 (-0.38)	64.75 (-0.44)	64.94 (-0.25)	64.31 (-0.88)	**65.13** (-0.06)
FGVCAircraft	Rob.	0.00	37.31 (+37.31)	**37.38** (+37.38)	15.88 (+15.88)	0.19 (+0.19)	29.81 (+29.81)
	Acc.	13.94	13.44 (-0.50)	13.25 (-0.69)	13.69 (-0.25)	13.69 (-0.25)	**13.81** (-0.13)
Avg.	Rob.	0.00	76.63 (+76.63)	**76.65** (+76.65)	33.08 (+33.08)	2.74 (+2.74)	66.36 (+66.36)
	Acc.	65.37	64.91 (-0.46)	64.79 (-0.58)	65.21 (-0.16)	64.95 (-0.42)	**65.34** (-0.03)

Table 7. Performance of ATAC with different augmentation settings under a 10-step PGD attack with $\epsilon = 4/255$, evaluated on 1,600 randomly sampled images from 5 datasets for each augmentation setting.

Algorithm 1: Adaptive ATAC Attack

Input: image $x \in [0,1]^{C\times H\times W}$
label y
CLIP image encoder E_I
text embeddings $\{t_i\}_{i=1}^{k}$
attack budget ϵ
attack step size γ
strategy weight λ
optimization steps T
gating temperature $\mathcal{K}$
ATAC augmentation functions $\{\mathcal{A}_i\}_{i=1}^{n}$
ATAC correction step size α
ATAC gating threshold τ^*
attack strategy strategy $\in \{\text{avoid}, \text{lure}\}$
Output: Adversarial perturbation $\delta^\star$ with $\|\delta^\star\|_\infty \le \epsilon$.

$\delta \sim \mathcal{U}(-\epsilon, +\epsilon)$
for $t = 1 \dots T$ **do**
  $x_a = x + \delta$

  `// ATAC`
  $f_x \leftarrow E_I(x_a)$
  $x_1, ..., x_n \leftarrow \mathcal{A}_i(x_a), ..., \mathcal{A}_i(x_a)$
  $f_{x_1}, ..., f_{x_n} \leftarrow E_I(x_1), ..., E_I(x_n)$
  $d_1, ..., d_n \leftarrow f_x - f_{x_1}, ..., f_x - f_{x_n}$
  $\bar{d} \leftarrow \frac{1}{n}\sum_{i=1}^{n} d_i$
  $\tau \leftarrow \frac{1}{n}\sum_{i=1}^{n} \cos(d_i, \bar{d})$

  `// Soft correction`
  $g \leftarrow \sigma(\mathcal{K} \cdot (\tau - \tau^*))$
  $f^* \leftarrow f_x + \alpha \cdot g \cdot \bar{d}$

  `// Strategy-dependent update`
  if strategy = avoid **then**
    logits $\leftarrow$ pred$(f_x, \{t_i\}_{i=1}^{k})$
    $l = \mathcal{L}(\text{logits}, y) - \lambda \cdot \tau$
  else
    logits $\leftarrow$ pred$(f^*, \{t_i\}_{i=1}^{k})$
    $l = \mathcal{L}(\text{logits}, y) + \lambda \cdot \tau$
  $\delta \leftarrow \prod_S(\delta + \gamma \cdot \text{sign}(\nabla_\delta l))$
$\delta^* \leftarrow \delta$
return δ^*.

Algorithm 2: Adaptive TTC Attack

Input: image $x \in [0,1]^{C\times H\times W}$
label y
CLIP image encoder E_I
text embeddings $\{t_i\}_{i=1}^{k}$
attack budget ϵ
attack step size γ
strategy weight λ
optimization steps T
gating temperature $\mathcal{K}$
TTC counterattack budget ϵ_{ttc}
TTC counterattack step size η
TTC gating threshold τ_{thresh}
TTC noise budget ϵ_τ
attack strategy strategy $\in \{\text{avoid}, \text{lure}\}$
Output: Adversarial perturbation $\delta^\star$ with $\|\delta^\star\|_\infty \le \epsilon$.

$\delta \sim \mathcal{U}(-\epsilon, +\epsilon)$
for $t = 1, ..., T$ **do**
  $x_a \leftarrow x + \delta$
  $f_x \leftarrow E_I(x_a)$

  `// TTC step`
  $\delta_{ttc} \sim \mathcal{U}(-\epsilon, +\epsilon)$
  $f_{x_{ttc}} \leftarrow E_I(x_a + \delta_{ttc})$
  $\delta_{ttc} \leftarrow \prod_S(\delta_{ttc} + \eta \cdot \text{sign}(\nabla_{\delta_{ttc}} \|f_x - f_{x_{ttc}}\|_2))$

  `// Calculating` $\hat{\tau}$ `via EOT for thresholding`
  $\hat{\tau} \leftarrow 0$
  for *j = 1, ..., K* **do**
    $n_i \sim \mathcal{U}(-\epsilon_\tau, +\epsilon_\tau)$
    $\hat{\tau} \leftarrow \hat{\tau} + \frac{1}{K} \cdot \frac{\|E_I(x_a+n) - f_x\|_2}{\|f_x\|_2}$

  `// Soft counterattack`
  $g \leftarrow \sigma(\mathcal{K} \cdot (\tau_{thresh} - \hat{\tau}))$
  $x^* \leftarrow x_a + g \cdot \delta_{ttc}$

  `// Strategy-dependent update`
  if strategy = avoid **then**
    logits $\leftarrow$ pred$(f_x, \{t_i\}_{i=1}^{k})$
    $l \leftarrow \mathcal{L}(\text{logits}, y) + \lambda \cdot \hat{\tau}$
  else
    logits $\leftarrow$ pred$(E_I(x^*), \{t_i\}_{i=1}^{k})$
    $l \leftarrow \mathcal{L}(\text{logits}, y) - \lambda \cdot \hat{\tau}$
  $\delta \leftarrow \prod_S(\delta + \gamma \cdot \text{sign}(\nabla_\delta l))$
$\delta^* \leftarrow \delta$
return δ^*.