CLIP is Strong Enough to Fight Back: Test-time Counterattacks towards Zero-shot Adversarial Robustness of CLIP

Supplementary Material

7. Analysis of τ

In the paper, we define a stochastic variable τ (Eq. (4)). In this section, we provide more results of τ and theoretical analyses. We report the values of τ for 16 datasets in Fig. 6. As can be seen from Fig. 6, when the random noise added onto the image is small, the resultant L_2 drift of the adversarial images in the embedding space is unusually small, indicating that they are trapped in their toxic local surroundings induced by the adversaries that seek to maximise the classification loss of CLIP. This behaviour is termed as 'false stability' in the main paper. When the strength of the random noise is sufficiently large, the L_2 drift of adversarial images is disproportionately enlarged. In contrast, the values of τ increase more steadily for clean images, as the noise strength ϵ_{random} increases, without showing disproportionate changes. Below we theoretically analyse the behaviour of 'false stability' of adversarial images.

7.1. Theoretical Analysis

Given a pre-trained vision encoder f_θ, a natural (unattacked) image $x \in \mathcal{R}^{C\times W\times H}$, and an adversarial image x' that is manipulated to maximise the classification loss of CLIP:

$$x' = \arg\max_{x_s} L(f_\theta(x_s), t_c), \quad s.t. \, \|x_s - x\|_\infty \leq \epsilon \tag{7}$$

the resultant embedding $f_\theta(x+n)$ when a small random noise $n \in \mathcal{R}^{C\times W\times H} \sim U(-\epsilon_{random}, \epsilon_{random})$ is imposed can be written as the Taylor expansion of f at x:

$$f_\theta(x+n) = f_\theta(x) + J_f(x)\cdot n + \frac{1}{2} n^T \cdot H_f(x) \cdot n + \cdots \tag{8}$$

where $J_f(x)$ and $H_f(x)$ are the Jacobian matrix and Hessian matrices of f at x, respectively, assuming that f is smooth around x. Provided that the random noise n is small, the above embedding can be approximated by the first-order expansion:

$$f_\theta(x+n) \approx f_\theta(x) + J_f(x)\cdot n \tag{9}$$

Therefore, the L_2 drift induced by n can be written as:

$$\begin{aligned} \| f_\theta(x+n) - f_\theta(x) \| &\approx \| J_f(x)\cdot n \| \\ &= \left(\sum_{j=1}^{d} \left(\sum_{i=1}^{N} \frac{\partial f_j}{\partial x_i} n_i \right)^2 \right)^{\frac{1}{2}} \end{aligned} \tag{10}$$

where d is the latent space dimensionality of CLIP, and $N = C \times W \times H$ is the pixel space dimensionality.

When Eq. (7) is computed by gradient-based methods such as PGD [6], x' is obtained through gradient ascent in the direction that increases the classification loss L:

$$\frac{\partial}{\partial x} L(f_\theta(x), t_c) = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial x} = \frac{\partial L}{\partial f} \cdot J_f(x) \tag{11}$$

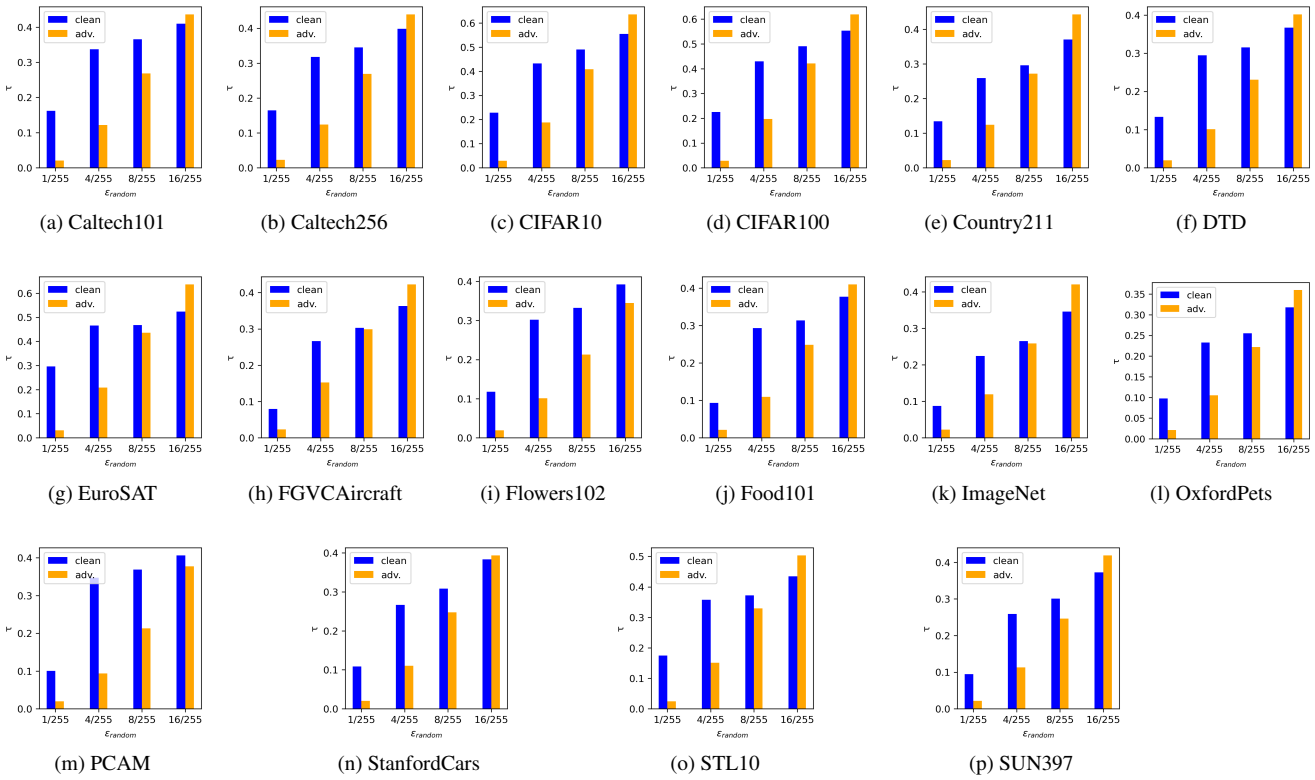
As such, the approximation of Eq. (7) can be seen as constantly searching the pixel space for the trajectory starting from x that causes the steepest ascent of L, i.e., the strongest activation of $J_f(x)$, within a limited number of steps. Since x' is the approximation result of Eq. (7), it lies in the trajectory where $J_f(x)$ is the most activated, and is therefore insensitive to a random noise n, which is statistically isotropic in the pixel space with a tiny component that lies in the direction of $\frac{\partial L}{\partial f} \cdot J_f(x)|_{x=x'}$. In contrast, a clean image x without being manipulated based on $J_f(x)$ does not show unusually strong activations in any direction, and can be more activated by an isotropic noise n. Therefore, $\|J_f(x)\cdot n\| > \|J_f(x')\cdot n\|$ holds when n is a small random noise in the pixel space, rendering the adversarial image x' 'falsely stable'.

8. More Results on Adversarial Robustness

In this section, we provide more complete results on adversarial robustness.

8.1. Robustness under CW attacks

Following previous studies [32, 50], we further test adversarial robustness of our test-time counterattack paradigm under CW attack [6], with the attack budget at $\epsilon_a = 1/255$. Tab. 4 reports the full table of results. It can be seen that for CW attacks, our TTC paradigm can still achieve stable robustness gains across 16 datasets. RN and TTE do not degrade accuracy on clean images since they do not counter the potential adversary by perturbing test images. Similarly to when tested under PGD attacks, TTE does not provide stable robustness. Compared to *Anti-adversary* and *HD*, which optimise a perturbation based on some objective, our TTC retains the best clean accuracy while significantly improving robustness. This shows that our paradigm can also be employed in test time to defend CLIP against other attack methods that maximise the classification loss of CLIP.



(a) Caltech101 (b) Caltech256 (c) CIFAR10 (d) CIFAR100 (e) Country211 (f) DTD

(g) EuroSAT (h) FGVCAircraft (i) Flowers102 (j) Food101 (k) ImageNet (l) OxfordPets

(m) PCAM (n) StanfordCars (o) STL10 (p) SUN397

Figure 6. Values of τ on clean and adversarial images ($\epsilon_a = 1/255$) across 16 datasets. For each dataset, we randomly sample 100 images and report the average values.

8.2. Robustness under $\epsilon_a = 4/255$

Tab. 5 reports the full table of robustness under 10-step PGD attack with the attack budget being $\epsilon_a = 4/255$. It can be seen that TTC achieves consistent and stable robustness gains across 16 datasets. *Anti-adversary* [1] and *HD* [52] bring little to no robustness under a high attack strength at $\epsilon_a = 4/255$. RN and TTE [38] perform best in terms of accuracy on clean images, which is understandable because they do not optimise any perturbation to counter the adversary. RN does not provide any robustness, showing that an additive random noise in the pixel space as large as the attack budget is not able to counteract the false stability of adversarial images. TTE [38] improves robustness of CLIP against PGD attacks with a high strength $\epsilon_a = 4/255$ to some extent. However, the robustness gain is unstable, as indicated by the high standard deviation of robust accuracy across different runs. For TTC, the number of steps N is increased to 5 for more effective counterattacks in this setting, which reduces the average clean accuracy by 5.52, compared to the original CLIP model. This trade-off is still reasonable given the consistent robustness gains.

9. Trade-off between Clean Accuracy and Robustness

Although we show that performing counterattacks at test time brings stable robustness gains, there is a reduction in clean accuracy, as can be seen in Tab. 1 and Tab. 5. Previous studies extensive analyse the trade-off between accuracy and robustness. TRADES [58] theoretically analyse such trade-off and propose to control it with a hyperparameter. In this work, the user-defined hyperparameter τ_{thres} adjusts the trade-off by determining whether a test image is 'falsely stable' or not, with a higher value benefiting robustness at a cost of clean accuracy. In our experiments, the reduction in clean accuracy is larger on some datasets such as ImageNet than others. This can be attributed to the different τ behaviour of clean images on various datasets. From Fig. 6, it can be seen that clean images of ImageNet exhibit lower values of τ when a random noise at the strength of $4/255$ is imposed. This causes more clean images of ImageNet to be considered as 'falsely stable' when the same τ_{thres} of 0.2 is employed, subjecting them to unnecessary counterattacking. In the future, we aim to further mitigate such trade-off by adaptively configuring τ_{thres} for downstream data distributions.

(%)		CLIP	Adversarial Finetuning				Test-time Defence					Δ
			CLIP-FT	TeCoA	PMG-AFT	FARE	RN	TTE	Anti-adv	HD	TTC (ours)	
TinyImageNet	Rob.	0.36	1.06	48.00	43.79	27.71	0.57±0.02	19.40±4.08	5.48±0.05	3.70±0.12	19.75±0.38	+19.39
	Acc.	57.64	77.06	70.86	66.85	73.63	51.85±0.04	56.73±0.22	52.76±0.16	52.49±0.12	51.85±0.04	-5.79
CIFAR10	Rob.	0.87	0.94	33.27	39.50	20.6	2.05±0.05	**40.01**±6.25	12.53±0.01	14.79±0.10	29.04±0.02	+28.17
	Acc.	85.12	84.90	64.61	70.69	74.44	81.18±0.07	**84.74**±0.40	83.52±0.09	78.64±0.02	81.18±0.07	-3.94
CIFAR100	Rob.	0.29	0.39	18.27	20.83	11.67	0.63±0.06	**18.73**±3.87	6.56±0.23	3.04±0.04	14.38±0.23	+14.09
	Acc.	57.14	59.51	35.96	40.32	46.67	56.34±0.20	**58.61**±0.25	53.95±0.15	53.50±0.02	56.34±0.20	-0.80
STL10	Rob.	12.23	9.95	69.73	72.39	59.60	17.20±0.15	**78.64**±3.91	38.66±0.17	37.73±0.22	76.40±0.16	+64.17
	Acc.	96.40	94.49	87.40	88.56	91.72	95.85±0.04	**96.26**±0.04	95.45±0.08	89.54±0.05	95.85±0.04	-0.55
ImageNet	Rob.	1.46	1.27	18.28	19.42	27.71	2.21±0.00	29.77±4.19	9.37±0.05	7.46±0.05	**36.01**±0.15	+34.55
	Acc.	59.69	54.24	34.89	36.12	48.79	59.34±0.06	**60.02**±0.13	54.27±0.14	55.06±0.05	49.39±0.00	-10.30
Caltech101	Rob.	20.88	15.95	56.23	61.58	54.86	25.89±0.11	**69.44**±3.09	41.47±0.02	36.26±0.08	66.17±0.31	+45.29
	Acc.	85.66	83.63	71.68	75.45	80.95	**86.61**±0.10	85.84±0.09	84.02±0.10	83.00±0.07	86.53±0.07	+0.87
Caltech256	Rob.	9.69	7.24	42.63	44.55	39.58	13.11±0.05	**59.81**±3.97	27.17±0.07	24.54±0.09	58.79±0.07	+49.10
	Acc.	81.72	78.53	61.14	62.24	73.32	81.25±0.03	**82.48**±0.08	79.38±0.12	79.38±0.05	79.66±0.04	-2.06
OxfordPets	Rob.	1.64	1.14	37.91	39.28	33.85	3.11±0.04	51.12±6.98	22.99±0.52	13.84±0.27	**57.15**±0.61	+55.51
	Acc.	87.44	84.14	62.12	65.88	79.37	87.41±0.12	**88.13**±0.13	80.62±0.35	80.64±0.15	83.35±0.21	-4.09
Flowers102	Rob.	1.35	0.80	21.13	21.34	17.25	2.13±0.06	34.97±4.25	8.06±0.07	8.51±0.04	**36.84**±0.13	+35.49
	Acc.	65.46	53.37	36.80	37.00	47.98	64.62±0.19	**65.20**±0.23	62.66±0.14	57.79±0.08	64.16±0.19	-1.30
FGVCAircraft	Rob.	0.00	0.00	2.25	1.86	1.35	0.00±0.00	5.15±1.25	0.83±0.11	0.97±0.06	**12.41**±0.32	+12.41
	Acc.	20.10	14.04	5.31	5.55	10.86	19.25±0.18	**20.18**±0.35	15.88±0.23	16.18±0.21	18.00±0.16	-2.10
StanfordCars	Rob.	2.38	2.04	8.74	10.53	9.14	2.44±0.02	21.19±3.41	4.76±0.18	5.11±0.05	**30.38**±0.12	+28.00
	Acc.	52.02	42.11	20.91	25.44	38.68	52.14±0.09	**52.73**±0.31	36.21±0.27	43.60±0.05	48.16±0.16	-3.86
SUN397	Rob.	1.75	1.48	18.36	20.39	15.73	2.48±0.03	29.37±4.05	8.85±0.01	7.90±0.03	**39.44**±0.07	+37.69
	Acc.	58.50	55.73	36.69	37.98	52.42	**59.69**±0.06	59.12±0.08	56.00±0.04	54.07±0.01	55.13±0.06	-3.37
Country211	Rob.	0.08	0.05	1.46	1.74	0.92	0.15±0.02	3.00±0.74	0.72±0.05	0.75±0.02	**6.17**±0.11	+6.09
	Acc.	15.25	12.07	4.75	4.64	9.26	**14.80**±0.02	14.66±0.14	11.58±0.12	11.98±0.02	13.08±0.05	-2.17
Food101	Rob.	1.09	0.55	12.87	16.57	12.93	1.92±0.04	44.61±6.42	15.03±0.11	9.77±0.06	**54.65**±0.13	+53.56
	Acc.	83.88	64.86	29.98	36.61	55.31	83.44±0.04	**83.96**±0.01	75.81±0.22	81.02±0.05	82.18±0.02	-1.70
EuroSAT	Rob.	0.03	0.03	11.66	11.94	10.66	0.16±0.00	6.44±1.74	2.57±0.08	3.47±0.17	**12.69**±0.07	+12.66
	Acc.	42.59	27.64	16.58	18.53	21.88	**53.24**±0.09	44.38±1.62	36.78±0.18	40.12±0.13	**53.24**±0.09	+10.65
DTD	Rob.	2.87	2.77	16.28	13.72	14.36	3.46±0.04	22.62±2.06	6.06±0.04	10.11±0.16	**27.39**±1.07	+24.52
	Acc.	40.64	36.49	25.16	21.76	32.07	37.96±0.13	**41.35**±0.29	38.92±0.22	35.25±0.22	36.98±0.21	-3.66
PCAM	Rob.	0.10	1.10	48.29	46.36	16.41	0.44±0.02	10.70±3.25	5.07±0.02	46.92±0.10	**52.86**±0.06	+52.76
	Acc.	52.02	47.21	49.96	50.03	52.54	**52.73**±0.07	50.92±0.04	52.49±0.02	50.35±0.01	**52.73**±0.07	+0.71
Avg.	Rob.	3.54	2.86	26.09	27.62	20.86	4.84±0.01	32.85±3.70	13.17±0.04	14.45±0.03	**38.17**±0.09	+34.63
	Acc.	61.51	55.80	40.25	42.30	51.02	61.61±0.03	**61.79**±0.13	57.35±0.03	56.88±0.02	59.75±0.06	-1.76

Table 4. Classification accuracy (%) on both adversarial images (Rob.) under 10-step CW attack [6] at $\epsilon_a = 1/255$ and clean images (Acc.) across 16 datasets. Weights and gradients of the deployed model are assumed to be known to the threat model. Comparison is made among our paradigm and test-time defences adapted from existing adversarial studies, with finetuning-based models implemented as a reference. We report the mean and standard deviation for test-time methods over 3 runs. The last column reports the gains w.r.t. original CLIP without any finetuning or test-time operations.

10. Pitfalls of Adversarial Finetuning

In the main paper, we find that although TTC can further improve robustness of adversarially finetuned CLIP models at test time (Tab. 3), the robustness gains are less obvious compared to the original CLIP. We also find that employing TTC on unsupervised adversarial finetuning method FARE [44] achieves greater gains compared to when employing TTC on TeCoA [32] and PMG-AFT [50], which are supervised adversarially finetuned CLIP models. Since our TTC paradigm is based on the expressiveness of the pre-trained vision encoder f_θ, we investigate this behaviour from the perspective of f_θ. Through analysis of randomly sampled images, we find that adversarial finetuning significantly reduces the sensitivity of f_θ to nuanced variations in the pixel space. We study the values of τ of different adversarially finetuned vision encoders when a random noise is imposed on clean images and report the results in Fig. 7. As can be seen from the figure, adversarial finetuning reduces the sensitivity of f_θ to pixel-level variations by orders of magnitude, which we believe is the key mechanism through which the adversarially finetuned models of CLIP achieve robustness against adversaries. Regular finetuning of CLIP

(%)		CLIP	Adversarial Finetuning							Test-time Defence					Δ
			CLIP-FT	TeCoA^1	TeCoA^4	PMG-AFT^1	PMG-AFT^4	FARE^1	FARE^4	RN	TTE	Anti-adv	HD	TTC (ours)	
TinyImageNet	Rob.	0.00	2.19	4.87	10.12	4.39	9.59	0.29	1.24	0.00±0.00	1.77±1.28	0.09±0.01	0.01±0.00	6.75±0.21	+6.75
	Acc.	57.64	77.06	70.86	63.84	66.85	59.77	73.63	70.69	51.85±0.04	56.73±0.22	52.62±0.20	51.07±0.09	51.85±0.04	-5.79
CIFAR10	Rob.	0.43	2.75	7.69	11.7	10.20	15.59	1.94	5.42	0.00±0.00	3.47±2.77	0.32±0.02	1.67±0.08	**28.51**±0.36	+28.08
	Acc.	85.12	84.90	64.61	65.15	70.69	71.45	74.44	78.46	81.18±0.07	**84.74**±0.40	83.44±0.07	78.23±0.16	81.18±0.07	-3.94
CIFAR100	Rob.	0.05	0.67	6.54	9.25	7.60	10.80	2.64	4.54	0.00±0.00	1.37±0.96	0.22±0.03	0.00±0.00	**9.06**±0.11	+9.01
	Acc.	57.14	59.51	35.96	36.30	40.32	41.51	46.67	47.38	56.34±0.20	**58.61**±0.25	53.96±0.17	52.86±0.16	56.34±0.20	-0.80
STL10	Rob.	0.16	3.75	24.80	31.83	28.49	35.40	9.99	17.59	0.06±0.01	32.56±11.76	2.25±0.10	3.39±0.12	**52.40**±0.34	+52.24
	Acc.	96.40	94.49	87.40	81.69	88.56	84.35	91.72	89.11	95.85±0.04	**96.26**±0.04	95.47±0.06	89.50±0.07	95.83±0.03	-0.57
ImageNet	Rob.	0.00	0.07	1.65	3.00	2.07	3.34	0.16	0.65	0.00±0.00	6.31±3.32	0.15±0.00	0.01±0.00	**12.68**±0.03	+12.68
	Acc.	59.69	54.24	34.89	27.76	36.12	28.51	48.79	40.48	59.34±0.06	**60.02**±0.13	54.29±0.07	54.54±0.05	34.00±0.06	-25.69
Caltech101	Rob.	0.59	4.81	15.75	21.00	19.48	25.03	5.15	10.13	0.68±0.02	30.19±7.92	3.14±0.07	1.27±0.03	**36.66**±0.25	+36.07
	Acc.	85.66	83.63	71.68	64.41	75.45	69.06	80.95	76.58	**86.61**±0.10	85.84±0.09	83.99±0.07	82.33±0.04	86.15±0.08	+0.49
Caltech256	Rob.	0.12	1.41	8.29	11.76	10.65	13.68	2.18	5.09	0.16±0.00	23.23±7.77	1.44±0.03	0.34±0.02	**27.25**±0.08	+27.13
	Acc.	81.72	78.53	61.14	52.05	62.24	53.32	73.32	67.22	81.25±0.03	**82.48**±0.08	79.40±0.07	79.12±0.01	76.59±0.12	-5.13
OxfordPets	Rob.	0.00	1.66	0.90	3.71	1.74	5.10	0.19	0.30	0.00±0.00	3.18±2.94	0.10±0.04	0.00±0.00	**24.64**±0.53	+24.64
	Acc.	87.44	84.14	62.12	53.94	65.88	56.66	79.37	70.10	87.41±0.12	**88.13**±0.13	80.53±0.17	80.91±0.05	64.70±0.33	-22.74
Flowers102	Rob.	0.00	0.13	1.87	3.81	2.57	4.26	0.03	0.62	0.00±0.00	3.52±2.51	0.05±0.02	0.00±0.00	**13.60**±0.33	+13.60
	Acc.	65.46	53.37	36.80	27.78	37.00	28.88	47.98	41.01	64.62±0.19	**65.20**±0.23	62.80±0.02	58.22±0.12	63.24±0.21	-2.22
FGVCAircraft	Rob.	0.00	0.00	0.03	0.12	0.03	0.06	0.00	0.03	0.00±0.00	0.43±0.43	0.00±0.00	0.00±0.00	**6.40**±0.38	+6.40
	Acc.	20.10	14.04	5.31	3.51	5.55	3.24	10.86	7.77	19.25±0.18	**20.18**±0.35	15.64±0.17	16.36±0.03	15.99±0.04	-4.11
StanfordCars	Rob.	0.00	0.00	0.15	0.41	0.15	0.40	0.01	0.04	0.00±0.00	1.46±1.21	0.00±0.00	0.00±0.00	**12.84**±0.20	+12.84
	Acc.	52.02	42.11	20.91	15.18	25.44	16.79	38.68	32.09	52.14±0.09	**52.73**±0.31	36.14±0.30	44.28±0.02	41.52±0.15	-10.50
SUN397	Rob.	0.00	0.02	1.30	2.31	1.90	3.24	0.13	0.57	0.00±0.00	5.95±3.39	0.11±0.00	0.00±0.00	**13.43**±0.08	+13.43
	Acc.	58.50	55.73	36.69	28.16	37.98	29.93	52.42	43.57	**59.69**±0.06	59.12±0.08	55.99±0.04	53.17±0.02	46.68±0.02	-11.82
Country211	Rob.	0.00	0.00	0.05	0.19	0.12	0.24	0.00	0.02	0.00±0.00	0.24±0.15	0.00±0.00	0.00±0.00	**2.44**±0.15	+2.44
	Acc.	15.25	12.07	4.75	3.66	4.64	3.34	9.26	6.58	**14.80**±0.02	14.66±0.14	11.60±0.08	11.72±0.07	11.99±0.01	-3.26
Food101	Rob.	0.00	0.04	0.56	1.35	1.03	2.12	0.06	0.24	0.00±0.00	5.31±4.09	0.07±0.02	0.01±0.00	**17.89**±0.13	+17.89
	Acc.	83.88	64.86	29.98	21.90	36.61	27.97	55.31	41.98	83.44±0.04	**83.96**±0.01	75.95±0.17	80.30±0.05	80.00±0.07	-3.88
EuroSAT	Rob.	0.00	0.00	9.77	10.71	9.61	10.36	0.00	7.34	0.00±0.00	0.11±0.09	0.03±0.02	0.20±0.02	**13.57**±0.12	+13.57
	Acc.	42.59	27.64	16.58	17.53	18.53	19.19	21.88	18.22	**53.24**±0.09	44.38±1.62	36.81±0.12	39.08±0.06	**53.24**±0.09	+10.65
DTD	Rob.	0.11	0.00	4.20	5.16	4.31	5.21	0.90	2.50	0.11±0.00	7.16±2.32	0.37±0.04	0.16±0.04	**11.40**±0.28	+11.29
	Acc.	40.64	36.49	25.16	20.11	21.76	17.29	32.07	28.03	37.96±0.13	**41.35**±0.29	38.55±0.12	34.89±0.35	35.69±0.08	-4.95
PCAM	Rob.	0.00	0.00	20.54	44.13	12.59	36.38	0.64	3.74	0.00±0.00	0.22±0.23	0.25±0.03	12.04±0.11	**47.39**±0.20	+47.39
	Acc.	52.02	47.21	49.96	49.98	50.03	49.80	52.54	50.17	**52.73**±0.07	50.92±0.04	52.61±0.07	50.38±0.04	**52.73**±0.07	+0.71
Avg.	Rob.	0.09	0.96	6.51	10.03	7.03	10.70	1.50	3.67	0.06±0.00	7.79±3.23	0.53±0.00	1.19±0.01	**20.63**±0.05	+20.54
	Acc.	61.51	55.80	40.25	35.57	42.30	37.58	51.02	46.17	61.61±0.03	**61.79**±0.13	57.32±0.03	56.62±0.02	55.99±0.06	-5.52

Table 5. Classification accuracy (%) on both adversarial images (Rob.) under 10-step PGD attack at $\epsilon_a = 4/255$ and clean images (Acc.) across 16 datasets. Weights and gradients of the deployed model are assumed to be known to the threat model. Comparison is made among our paradigm and test-time defences adapted from existing adversarial studies, with finetuning-based models implemented as a reference. The superscripts of the model names indicate the attack budget used for generating adversarial images in the phase of adversarial finetuning. We report the mean and standard deviation for test-time methods over 3 runs. The last column reports the gains w.r.t. original CLIP without any finetuning or test-time operations.

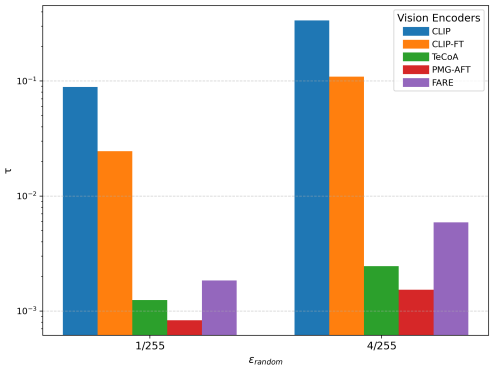


Figure 7. Average τ of different CLIP vision encoders on randomly sampled clean images across 16 datasets.

(CLIP-FT), i.e., finetuning the vision encoder with clean images on TinyImageNet, also reduces the perception sensitivity to some extent. Among adversarially finetuned models, FARE shows greater preservation of sensitivity compared to its supervised counterparts TeCoA and PMG-AFT, which explains the lower levels of adversarial robustness of FARE (Tab. 1) and better robustness gains when employing TTC on FARE at test time (Tab. 3). Although resilience to pixel-level variations translates to robustness of CLIP to imperceptible malicious perturbations, it causes the vision encoder to be less expressive. We argue that a fundamental difference between CLIP and non-foundational models is that CLIP has learned massive amounts of real-world knowledge, which should be taken into account in attempts aiming to enhance its robustness. We also recommend cautious use of adversarial finetuning as the only robustifying approach for CLIP and other large pre-trained models in general.

11. Effects of Other Hyperparameters

In the main paper, we find that the number of counterattack steps N is the crucial hyperparameter that greatly impacts robustness. In this section, we investigate the impact of the other two hyperparameters τ_{thres} (Eq. (4)) and β (Eq. (5)), which control the threshold of L_2 drift ratio and the ascending rate of weights across counterattack steps, respectively (Algorithm 1). We vary one hyperparameter at a time w.r.t. the default setting $\tau_{thres} = 0.2$ and $\beta = 2.0$. The counterattack budget ϵ_{ttc} and steps N are fixed to $\epsilon_{ttc} = 4/255$ and $N = 2$, respectively. We report the results in Tab. 6. It can be seen that both hyperparameters control the trade-off between the accuracy on clean images and adversarial images. When the threshold τ_{thres} is relatively small, the accuracy on clean images can be better retained, while the robustness gains are limited, since the values of τ for most clean and adversarial images are above the set threshold, which halts necessary counterattacks. Robustness increases as τ_{thres} is set higher, and reaches a plateau after $\tau_{thres} = 0.2$. Further increasing the threshold compromises accuracy on clean images. The impact of β is less obvious. In general, a larger β assigns higher weights to counterattack perturbations at later steps, thereby favouring robustness.

12. Adaptive Attacks

In the paper, we demonstrate that CLIP possesses the ability to defend itself from adversarial attacks that aim to maximise the classification loss of CLIP, assuming that such counterattacks by the end user are not known to the attacker. Here we provide a gradient-based method tailored to break our TTC. Our TTC paradigm can be written as $\varphi(x) = x + \delta^*(x)$, where x is a test image and δ^* is a function of x that induces the maximum L_2 drift of x in the embedding space of CLIP:

$$\delta^*(x) = \arg\max_{\delta} \|f_\theta(x+\delta) - f_\theta(x)\|,\ s.t.\ \|\delta\| \leq \epsilon_{ttc} \tag{12}$$

Therefore, the attacker may incorporate $\varphi(x)$ into the objective when crafting an adversarial image aiming to maximise the classification loss:

$$x' = \arg\max_{x_s} L(f_\theta(\varphi(x_s)), t_c),\ s.t.\ \|x_s - x\| \leq \epsilon_a \tag{13}$$

When employing gradient-based attack methods such as PGD to solve Eq. (13), the inner optimization of Eq. (12) can be approximated by a one-step update:

$$\begin{aligned} \varphi(x) &= x + \delta^*(x) \\ &\approx x + \delta^0 + \eta \nabla_\delta \|f_\theta(x+\delta^0) - f_\theta(x)\| \end{aligned} \tag{14}$$

where η is the step size for the counterattack and $\delta^0 \sim U(-\epsilon_{ttc}, \epsilon_{ttc})$ is a randomly initialised noise δ^0. Thus, the objective for generating the adversarial attack can be written as $L\left(f_\theta(x+\delta^0 + \eta\nabla_\delta \|f_\theta(x+\delta^0) - f_\theta(x)\|),\ t_c\right)$. By employing PGD to craft an adversary that maximises this objective, the attacker may break the counterattacks performed by the end user.

τ_{thres}	β	CIFAR10	CIFAR100	STL10	ImageNet	Caltech101	Caltech256	OxfordPets	Flower102	FGVCAircraft	StanfordCars	SUN397	Country211	Food101	EuroSAT	DTD	PCAM	Avg. Rob.	Avg. Acc.
0.2	2.0	28.75	14.31	76.70	38.41	65.78	60.11	57.87	39.14	13.77	33.01	41.52	7.09	57.84	12.19	27.32	52.85	39.17	59.75
0.05	2.0	2.07	0.69	16.35	2.77	19.14	11.83	3.35	2.75	0.00	0.37	2.35	0.12	1.51	0.14	5.74	3.72	4.56	61.63
0.1	2.0	2.15	0.88	26.21	18.95	32.96	28.28	32.35	25.13	2.19	16.78	17.06	2.35	29.31	0.67	17.66	37.25	18.14	61.46
0.15	2.0	7.62	4.96	56.45	33.06	55.43	50.48	52.77	36.82	9.33	30.33	34.07	5.48	53.22	5.66	25.21	48.50	31.84	61.00
0.25	2.0	44.20	22.29	83.09	40.18	69.32	63.45	58.93	40.01	14.97	33.42	43.64	7.73	58.63	18.16	28.46	55.49	42.62	57.31
0.3	2.0	50.10	25.94	84.86	40.66	70.43	64.48	59.12	40.14	15.30	33.44	44.30	7.89	58.87	21.19	28.88	56.65	43.89	54.18
0.35	2.0	51.97	27.07	85.49	40.83	70.76	64.94	59.23	40.15	15.45	33.48	44.49	7.96	58.95	22.64	29.10	57.17	44.35	50.67
0.4	2.0	52.36	27.51	85.56	40.91	70.89	65.10	59.25	40.15	15.51	33.48	44.56	7.99	58.99	23.44	29.15	57.40	44.52	47.75
0.2	0.5	27.08	13.25	74.58	33.53	63.50	57.44	48.24	32.90	10.98	27.75	36.45	5.71	51.96	12.11	24.95	41.20	35.10	60.24
0.2	1.0	28.01	13.84	75.97	36.39	64.94	59.12	53.80	36.27	12.72	30.95	39.35	6.48	55.60	12.44	26.65	48.03	37.54	60.00
0.2	1.5	28.42	14.02	76.46	37.73	65.54	59.81	56.55	38.12	13.50	32.28	40.80	6.90	57.18	12.49	27.45	51.25	38.66	59.85
0.2	2.5	28.82	14.13	76.81	38.77	65.89	60.25	58.54	39.76	13.71	33.44	41.83	7.28	58.10	12.55	27.77	53.98	39.48	59.72
0.2	3.0	28.95	14.15	76.91	38.95	66.03	60.34	58.90	40.06	13.92	33.62	42.07	7.36	58.25	12.54	27.87	54.50	39.65	59.70

Table 6. The Effects of hyperparameters τ_{thres} and β under 10-step PGD attack with $\epsilon_a = 1/255$. The counterattack budget and steps are fixed at $\epsilon_{ttc} = 4/255$ and $N = 2$, respectively. We report the robust accuracy for each dataset. The last column reports the average accuracy on clean images across 16 datasets.