TTP: Test-Time Padding for Adversarial Detection and Robust Adaptation on Vision-Language Models

Supplementary Material

A. Overview

This supplementary material presents comprehensive dataset details and additional experiments that were omitted from the main paper due to space constraints. These results further substantiate the effectiveness of our proposed Test-Time Padding (TTP).

B. Datasets

We summarize the content, number of categories and number of images for all datasets used in our experiments in Table 8.

Dataset	Description	# Classes	# Test
Caltech101	Object images	100	2,465
Pets	Pet images	37	3,669
Cars	Car images	196	8,041
Flower102	Flower images	102	2,463
Aircraft	Aircraft images	100	3,333
DTD	Describable textures dataset	47	1,692
EuroSAT	Sentinel-2 satellite images	10	8,100
UCF101	Human action images	101	3,783

Table 8. All datasets involved in our experiments.

C. Results of Different Thresholds

As demonstrated in Figure 4a of the main paper, with a padding size of 32, clean samples and adversarial examples exhibit significant differences in average cosine similarity when compared to their padded versions. Consequently, a preset threshold τ can be effectively employed to distinguish between them. In our experiments, we empirically set $\tau = 0.8$, which yielded robust cross-dataset and cross-model detection performance (as shown in Figure 2).

To ensure experimental integrity, we further investigated the sensitivity of detection accuracy to variations in the threshold value. As illustrated in Figure 5, our method consistently achieves optimal performance at $\tau = 0.8$ across all datasets. This indicates that our approach is dataset-agnostic, allowing for the use of a unified hyperparameter τ without the burden of extensive tuning, thereby underscoring the practical utility of TTP. Furthermore, detection accuracy degrades when the threshold deviates (either lower or higher) from this optimal value. This behavior is consistent with the cosine similarity distributions between clean samples and adversarial examples before and after padding, as illustrated in Figure 4a.

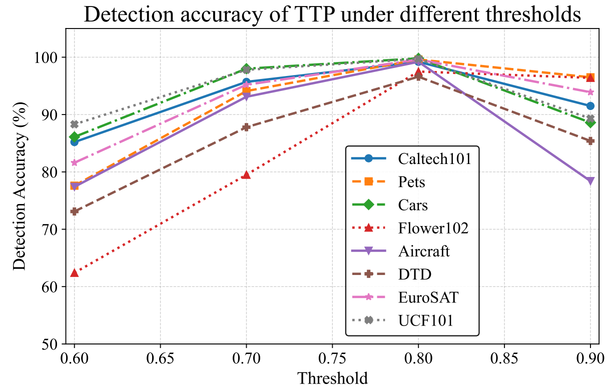

Figure 5. Detection accuracy of TTP with CLIP-ViT-B/32 ($\epsilon = 4.0$) under varying threshold values.

D. Robustness under Various Attacks

Due to space limitations in Section 4.2 of the main text, we restricted our comparison to the SOTA defense, R-TPT, under multiple adversarial attacks beyond PGD. To provide a comprehensive evaluation, we present a comparison against all baselines in this section. The results demonstrate that our proposed TTP consistently outperforms all competing baselines across the various attack scenarios.

Method	Flowers				DTD			
	CW	DF	FGSM	Avg.	CW	DF	FGSM	Avg.
CLIP [31]	0.8	0.4	4.8	2.0	2.3	7.6	13.4	7.8
TTC [46]	3.6	56.5	2.1	20.7	4.3	31.1	1.4	12.3
Ensemble	50.1	52.2	46.6	49.7	31.1	32.9	29.7	31.2
MTA [48]	34.5	35.4	36.6	35.5	23.6	23.5	23.9	23.7
R-TPT [33]	51.6	54.7	49.2	51.8	34.2	35.9	32.5	34.2
TTP (Ours)	**54.1**	**56.4**	**51.8**	**54.1**	**38.9**	**40.1**	**37.1**	**38.7**

Table 9. Adversarial accuracies (%) under CW, DeepFool (DF), and FGSM attacks on two fine-grained datasets. TTP achieves the best robustness across all attacks.

E. Performance on Complex Edges

We further evaluate the effectiveness of attention restoration on images with complex textures and edge patterns using the DTD dataset. As illustrated in Fig. 6, Row 1, trained padding can effectively restore the attention distribution even in the presence of dense edges and repetitive textures. This attention recovery translates into strong detection and defense performance on DTD, achieving over

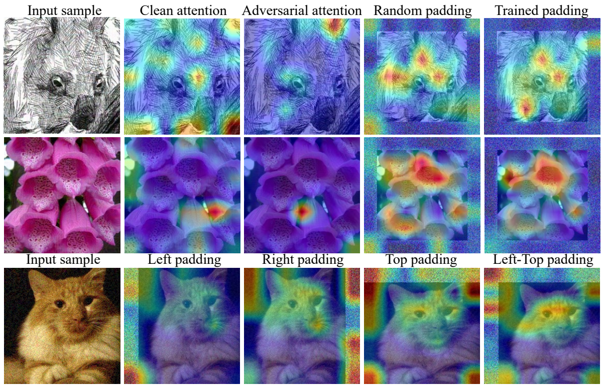

Figure 6. Visualization of **attention restoration.**

Input	Vanilla CLIP	R-TPT (SOTA)	TTP (Ours)
Clean samples	0.0057s	0.27s	0.0060s
Adversarial examples	0.0057s	0.27s	0.097s

Table 10. Average **inference time per sample** on an A100 GPU.

96% detection accuracy and SOTA robustness in our evaluation. These results suggest that the proposed padding strategy is not limited to simple scenes, but also remains effective for visually cluttered, edge-rich images.

F. Failure Case Analysis

Despite the overall effectiveness of our proposed detection method, we identify a failure pattern that warrants further discussion. As illustrated in Fig. 6 (Row 2), the detection tends to encounter challenges when the target object occupies nearly the entire canvas. In such scenarios, although adversarial perturbations succeed in significantly distorting the model's attention, the resulting focus remains spatially confined within the boundaries of the target object. Consequently, the extracted visual features are highly similar, hindering effective detection.

G. Asymmetric Padding

We additionally investigate asymmetric padding as a potential design choice. As shown in Fig. 6, Row 3, asymmetric padding is counterproductive. A plausible reason is that adding pixels only on one side introduces a strong positional bias, which shifts CLIP's attention toward the padded region. This attention shift interferes with feature extraction, ultimately causing a noticeable drop in recognition robustness. Based on these observations, we do not adopt asymmetric padding in our final pipeline.

H. Inference Latency

In realistic deployments, the majority of inputs are clean. In contrast to prior test-time defenses that apply adaptation to every sample (thereby paying unnecessary computation on clean inputs), our method performs only a lightweight detection step for clean samples and triggers adaptation only when needed. Moreover, our similarity-aware ensemble reduces the need for a large number of augmented views for final prediction, which further lowers inference time even on adversarial inputs. Empirically, on DTD (Tab. 10), our method is 45× faster than the SOTA R-TPT on clean samples (with latency comparable to vanilla CLIP) and 2.8× faster on adversarial samples. These results indicate that our approach improves robustness while maintaining practical inference efficiency.