ReMP-AD: Retrieval-enhanced Multi-modal Prompt Fusion for Few-Shot Industrial Visual Anomaly Detection

Supplementary Material

Table 7. Results under different layers $\hat{p}$ applied to VLPF on the VisA dataset in 4-shot setting

$\hat{p}$	I-AUC	I-F1
6 12 18 24	93.7	89.3
12 18 24	**94.1**	**89.9**
18 24	93.9	89.3
24	93.9	89.5

Table 8. Results using various values of w_1,w_2 applied to ICTR on VisA dataset under the 4-shot setting. AVG denotes the average of P-F1 and I-F1.

w_1-w_2	P-F1	I-F1	AVG
1.2-0.5	47.2	**90.0**	68.6
1.2-0.8	48.0	89.9	**68.9**
1.2-1.0	**48.1**	89.5	68.8
1.5-0.5	47.1	89.8	68.5
1.5-0.8	47.8	89.7	68.8
1.5-1.0	47.9	89.7	68.8
2.0-0.5	47.0	89.1	68.0
2.0-0.8	47.6	89.1	68.4
2.0-1.0	47.8	89.0	68.4

Table 9. Results using various values of λ applied to ICTR on VisA dataset under the 4-shot setting. AVG denotes the average of P-AUC, P-F1, I-AUC, and I-F1.

λ	P-AUC	P-F1	I-AUC	I-F1	AVG
0.4	97.0	43.2	93.1	88.5	80.5
0.5	97.3	44.4	94.0	89.3	81.2
0.6	97.4	45.7	**94.5**	**90.2**	82.0
0.7	97.7	**47.9**	94.2	90.0	**82.4**
0.8	**98.0**	46.3	92.5	88.0	81.2
0.9	97.3	42.0	90.7	86.4	79.1
0.64	97.3	46.9	**94.7**	90.4	82.3
0.66	97.6	47.3	94.4	89.9	82.3
0.68	97.5	47.8	94.4	89.8	82.4
0.70	97.7	47.9	94.2	90.0	82.4
0.72	97.8	**48.0**	94.1	**89.9**	**82.5**
0.74	**97.9**	47.8	93.5	89.2	82.1

Table 10. Comparison of results on MPDD and Real-IAD datasets. Metrics are P-AUC/I-AUC.

Dataset	Method	1-shot	2-shot	4-shot
MPDD	APRIL-GAN	**97.4**/76.5	97.5/76.7	97.6/77.9
	PromptAD	96.2/**80.7**	97.2/**85.3**	97.3/**87.2**
	ReMP-AD	**97.4/77.6**	**98.0**/83.0	**98.2**/86.2
Real-IAD	APRIL-GAN	**95.6/74.6**	96.4/75.5	96.7/76.6
	PromptAD	91.3/66.8	93.8/73.2	95.1/76.8
	ReMP-AD	95.0/74.4	**96.7/77.2**	**97.5/79.8**

7. Hyper-parameter analysis

This section completes the effects of $\hat{p}$, w_1, w_2 and λ are evaluated to assess their contributions to the model's overall performance. To assess the effect of the attention layers $\hat{p}$ in the VLPF image encoder, Table 7 demonstrates the impact of the mask attention layer. The inclusion of the 6th layer, along with the exclusion of the 12th and 18th layers, results in a decline in performance. To evaluate the attention weights w_1 and w_2 in the VLPF image encoder, Table 8 illustrates their effects. Higher values of w_1 and lower values of w_2 lead to a decrease in performance. Furthermore, to achieve a balance between pixel-level and image-level results, the optimal values are $w_1 = 1.2$ and $w_2 = 0.8$. The impact of λ in GPR is evaluated in Table 9, which reports its performance across a range of 0.4 to 0.9, reaching a peak at $\lambda = 0.7$. To refine this range further, we measure λ within 0.64 to 0.74. For optimal balance between image-level and pixel-level performance, $\lambda = 0.72$ is selected as the best value.

8. More results on other benchmarks

In addition, we evaluate ReMP-AD in the few-shot setting on the MPDD [11] and Real-IAD [22] datasets. We reproduce the results of APRIL-GAN and PromptAD for comparison. As shown in Table 10, ReMP-AD achieves the best pixel-level performance on the MPDD dataset. On the Real-IAD dataset, ReMP-AD outperforms APRIL-GAN and PromptAD in the 2-shot and 4-shot settings.

9. Comparisons with identical backbones

For fair comparison with PromptAD, we report the parameters and results of ReMP-AD with ViT-B/16 and ViT-B/16+ in Table 11. ReMP-AD has slightly more parameters, but consistently outperforms PromptAD on both backbones, highlighting the generality of our method.

10. More visualization results

Fig.5 and Fig.6 illustrate the anomaly segmentation results of ReMP-AD on the MVTec-AD and VisA datasets in the 4-shot setting. The results indicate that ReMP-AD is able to identify anomaly regions across a diverse set of categories accurately.

11. Addressing limitations in related work

The GPR module retrieves prototypical patterns from reference examples to reduce background noise in reference samples. Meanwhile, the CTB module further picks up most match samples according to the relevance between the test sample and reference samples to reduce intra-class noise. On the other hand, VLPF unit vision principle and vision-textual principle to generate region-level prompts to guide transformer capturing more distinctive and relevant embedding of anomalies.

Table 11. Comparison of results with identical backbone on VisA dataset. Metrics are P-AUC/P-F1/I-AUC/I-F1.

Backbone	Method	Parameters	1-shot	2-shot	4-shot
ViT-B-16	APRIL-GAN	151.83M	95.5/32.3/81.7/81.2	95.6/32.0/83.9/81.5	95.7/34.0/84.9/81.6
	PromptAD	**149.62M**	93.8/26.8/82.5/**83.6**	95.4/31.0/81.1/83.1	95.6/32.8/83.4/81.3
	ReMP-AD(ours)	151.83M	**96.9/35.4/83.3**/82.0	**96.7/35.9/85.3/83.6**	**97.0/39.6/86.6/83.4**
ViT-B-16+	APRIL-GAN	211.39M	95.9/31.4/83.8/82.0	96.1/32.4/85.2/83.2	96.3/32.7/86.3/84.0
	PromptAD	**208.38M**	95.9/34.8/85.4/82.2	96.4/35.9/84.6/81.7	96.8/36.5/87.9/83.9
	ReMP-AD(ours)	211.39M	**96.4/36.5/85.4/83.0**	**96.3/38.8/87.6/83.9**	**97.1/41.4/89.9/85.7**

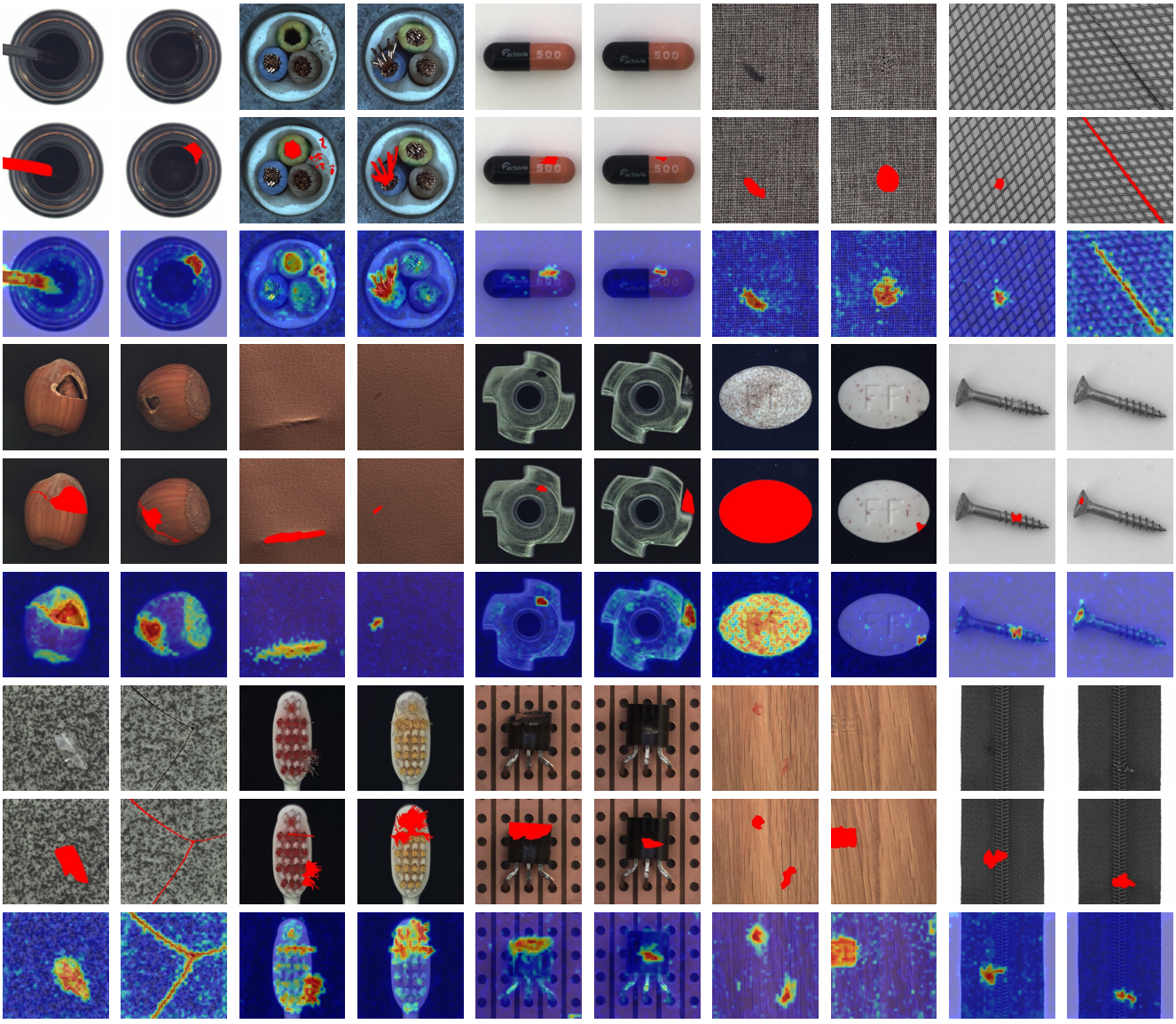

Figure 5. Anomaly segmentation results on the MVTec-AD dataset under the 4-shot anomaly detection setting. For each tuple, the images from top to bottom represent the anomaly image, ground truth, and predicted anomaly map.

Figure 6. Anomaly segmentation results on the VisA dataset under the 4-shot anomaly detection setting. For each tuple, the images from top to bottom represent the anomaly image, ground truth, and predicted anomaly map.