Learning Unified Reference Representation for Unsupervised Multi-class Anomaly Detection
Supplementary Material

1 Overview

This supplementary material consists of:

1. The ablation study of fusion approaches of MLKA and LCA. (Sec. 2)
2. The qualitative results on VisA dataset. (Sec. 3)
3. The ablation study of the quantity of learnable references. (Sec. 4)
4. The failure analysis. (Sec. 5)

2 Fusion of MLKA and LCA

In order to integrate the outputs of MLKA and LCA modules, we explored alternative fusion methods. In addition to employing the hyperparameter α to perform a weighted summation of their outputs, we conducted experiments by introducing learnable parameters to facilitate the weighted summation of the two modules. This modification involved transforming Equation (4) in the main text into

$$\begin{aligned} \mathcal{Z}_k = LN\,(\theta_1\ MLKA(\mathcal{Y}_{k-1}, \mathcal{R}_h) \\ + \theta_2\ LCA(\mathcal{Y}_{k-1}, \mathcal{R}_h))\,, \end{aligned} \tag{1}$$

where θ_1 and θ_2 are the parameters learned by the model. The experimental results are presented in Table 1, which demonstrate that the use of learnable weights increases the likelihood of the model learning shortcuts, potentially focusing more on MLKA rather than LCA. This suggests that by using the hyperparameter α, we can encourage the model to pay more attention to LCA, thereby compelling it to learn normal patterns and achieve better detection performance.

Table 1: Different fusion methods of MLKA and LCA on MVTec-AD. Metrics are presented in the form of Image-AUROC% / Pixel-AUROC%.

Fusion	w/ θ	w/ α
Metrics	98.1 / 98.2	**98.6** / **98.5**

3 Qualitative Results on VisA.

To further validate the superiority of our approach, we also conducted a qualitative analysis on the VisA [4] dataset. Figure 1 presents the visual detection results of UniAD [2] and our RLR on the VisA dataset. It is important to note that we were unable to reproduce the results of OmniAL [3] due to the unavailability of its source code, hence its results are not shown. Similar to the qualitative analysis conducted on the MVTec-AD [1] dataset, our method successfully detects samples that UniAD fails to detect, thus demonstrating the effectiveness of our approach.

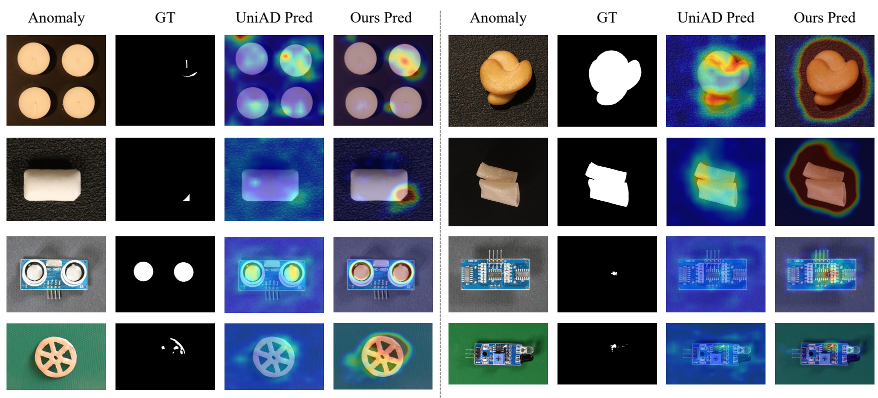

Fig. 1: Qualitative results on VisA. We visualize several anomalies (Anomaly) along with their corresponding Ground Truth (GT), the detection results of UniAD (UniAD Pred), and the detection results of our method (Ours Pred).

4 Multiple Learnable References

We conducted experimental analysis on the quantity of learnable references. Specifically, we performed calculations of attention with multiple references in the LCA and summed the outputs, which transformed $LCA(\mathcal{Y}_{k-1}, \mathcal{R}_h)$ in Equation (4) in the main text into $\sum_{i=1}^{K} LCA(\mathcal{Y}_{k-1}, \mathcal{R}_h^i)$, where K is the number of learnable references. Additionally, we imposed constraints on the diversity of references in the loss function. The experimental results are presented in Table 2, which indicates that this approach does not lead to performance improvement; instead, it incurs higher computational and storage resource consumption. Therefore, this paper only utilizes a single learnable reference representation ($K = 1$).

Table 2: The ablation study results of different K on MVTec-AD. Metrics are presented in the form of Image-AUROC% / Pixel-AUROC%.

$K =$	1	2	3
Metrics	**98.6** / **98.5**	**98.6** / **98.5**	**98.6** / 98.4

5 The Failure Analysis

We conducted a failure analysis on classes with low performance metrics on the MVTec-AD and VisA datasets. Specifically, in MVTec-AD, the Toothbrush class exhibits lower metrics due to the recognition of noise in the background as anomalies, resulting in some normal samples being misclassified as abnormal samples. This phenomenon is more likely to occur in feature reconstruction methods, as evidenced by similarly low metrics for this class in UniAD. Additionally, in the VisA dataset, the Capsules class exhibits lower metrics owing to the presence of random discolorations. Reconstruction models tend to confuse these discolorations with subtle anomalies, leading to poor performance of RLR and UniAD on this class.

References

1. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9592–9600 (2019)
2. You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., Le, X.: A unified model for multi-class anomaly detection. Advances in Neural Information Processing Systems **35**, 4571–4584 (2022)
3. Zhao, Y.: Omnial: A unified cnn framework for unsupervised anomaly localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3924–3933 (2023)
4. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European Conference on Computer Vision. pp. 392–408. Springer (2022)