Towards Accurate Unified Anomaly Segmentation (Supplementary Material)

This Supplementary Material contains the following parts: 1) Additional information details about sample-aware reweighting mechanisms and hyper-parameter settings in Appendix A; 2) Additional experiments, ablation studies, and visualization results in Appendix B.

A. Implementation Details

A.1. Sample-Aware Reweighting Mechanism

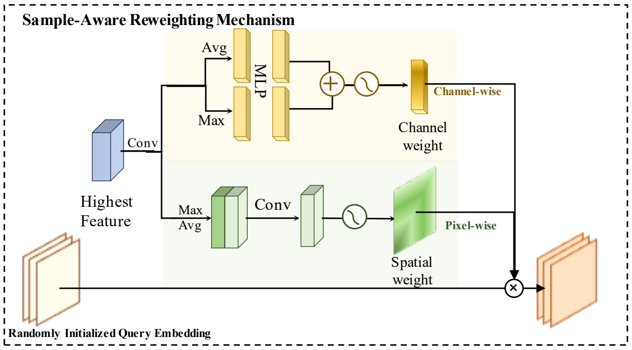

Figure 1. The illustration of the Sample-Aware Reweighting (SAR) Mechanism. Channel-wise attention and spatial-wise attention are computed based on the highest feature, which serves as the weights to reweight the randomly initialized query.

For One-for-All anomaly segmentation, a learnable query can be included to serve as the memory matrix, allowing for flexible memorization of class-agnostic semantics [4, 8]. However, features extracted from different samples, particularly those from distinct categories, often exhibit significant disparities in high-level characteristics. Employing a shared learnable query to memorize patterns across all categories can be suboptimal due to this variability. To address this issue, [4] introduces a Switching Mechanism with various codebooks and experts, although this approach is memory-consuming.

Instead, we take advantage of this variability in various features, incorporating a simple Sample-Aware Reweighting (SAR) Mechanism to initialize more sample-specific accurate prototypes as queries, as illustrated in Fig. 1. Specifically, the query $\mathbf{q}_0^{ori} \in \mathbb{R}^{H_K \times W_K \times C'}$ is randomly initialized, whose shape is the same as the highest patch embedding $\mathbf{h}_K$. We utilize channel weights $\mathbf{W}_c$ and spatial weights $\mathbf{W}_s$ in CBAM [7] to reweight query $\mathbf{q}_0^{ori}$ based on $\mathbf{h}_K$ with the richest semantic, as shown in Eq. (1).

$$\begin{aligned} \mathbf{W}_c &= \sigma(\text{MLP}(\text{AvgP}_s(\mathbf{h}_K) + \text{MaxP}_s(\mathbf{h}_K))), \\ \mathbf{W}_s &= \sigma(\text{Conv}(\text{cat}(\text{AvgP}_c(\mathbf{h}_K); \text{MaxP}_c(\mathbf{h}_K)))), \end{aligned} \tag{1}$$

where $\sigma(\cdot)$ is the activation function, $\text{AvgP}_s(\cdot), \text{MaxP}_s(\cdot)$ mean spatial-wise average-pooling and max-pooling, and $\text{AvgP}_c(\cdot), \text{MaxP}_c(\cdot)$ are channel-wise poolings.

The weights $\mathbf{W}_c, \mathbf{W}_s$ are channel-wisely and spatial-wisely multiplied to $\mathbf{q}_0^{ori}$, resulting in $\mathbf{q}_0 \in \mathbb{R}^{H_k \times W_k \times C'}$, which is the query of the first transformer layer.

A.2. Hyper-parameter Settings

The shapes of the four levels of features are $112 \times 112 \times 24$, $56 \times 56 \times 32$, $28 \times 28 \times 56$ and $14 \times 14 \times 160$, respectively. For the Gaussian filter, the kernel size is 3 and sigma equals 1. The patch size for each level is 8,4,2 and 1 accordingly. Inspired by [2], we combine two transformer components together in a transformer layer: a conventional spatial-wise transformer [6] and a channel-wise transformer that performs attention operations on channels after transposing the input. This dual design enables the model to simultaneously consider spatial-wise context and channel-wise semantics, enhancing the model's performance and robustness in dealing with intricate scenarios across multiple datasets. For both spatial- and channel-wise attention, the number of heads is 4, and the dimension of the feed-forward network is 2048. All the channel sizes in the convolution layers in MGG-CNN module are set to 256.

During evaluation, the threshold for segmentation is chosen based on the PR curve of each class to maximize the sum of precision and recall, and DSC is calculated sample-wisely. DSC for normal samples ($GT = 0$) is calculated based on the following rules: 1) if each pixel in the predicted segmentation mask equals 0, then the DSC for this image is 1; 2) if any pixel in the segmentation mask is wrongly predicted, then the DSC for this image is 0.

Category	AR(%)	Boundary	Img Recon.	Feature Recon.				
		PatchCore [5] CVPR2022	DRAEM [9] ICCV2021	DeSTSeg [10] CVPR2023	RD4AD [1] CVPR2022	UniAD [8] NIPS2022	HVQ-Trans [4] NIPS2023	**UniAS(Ours)**
Candle	0.81	13.21/ 0.27	14.02/ 0.27	32.08/ 0.27	19.80/ 0.00	20.77/ 0.34	19.47/ 5.38	19.09/ 0.36
Capsules	1.63	61.28/51.78	19.40/ 0.77	22.47/ 0.84	39.01/ 0.00	49.22/45.71	45.98/37.79	49.23/46.43
Cashew	5.06	54.81/38.27	1.38/ 1.84	51.89/47.04	37.30/ 0.01	42.91/ 8.02	59.50/36.01	52.43/38.64
Chewing Gum	2.54	45.53/51.74	43.68/45.78	61.59/73.50	62.09/32.22	57.97/61.22	47.70/55.71	62.79/54.32
Fryum	10.37	33.61/ 5.92	34.03/ 4.05	30.57/ 3.99	51.46/ 0.07	46.87/21.02	50.16/28.70	47.52/71.82
Macaroni1	0.22	1.06/50.01	14.82/ 0.07	0.62/50.24	22.10/ 0.05	9.62/ 0.03	7.29/50.74	14.73/50.78
Macaroni2	0.17	0.03/ 0.01	10.42/50.56	6.01/ 0.06	13.80/ 0.00	3.77/ 0.15	3.07/ 0.14	0.17/ 0.07
Pcb1	2.79	75.31/58.44	29.26/50.43	39.00/ 0.88	74.84/46.63	70.01/56.81	59.98/55.57	79.19/58.58
Pcb2	1.24	18.09/ 0.05	7.76/ 0.41	11.77/ 0.53	18.93/ 0.00	9.67/ 1.03	7.84/ 0.88	11.95/ 0.69
Pcb3	1.52	30.35/50.38	19.26/50.54	26.16/50.67	26.43/ 0.50	21.27/ 1.59	18.59/ 1.64	23.58/ 3.98
Pcb4	3.96	36.17/ 1.49	17.51/ 1.32	43.52/ 1.29	33.36/ 0.40	29.72/ 4.10	14.30/ 6.47	44.47/32.68
Pipe Fryum	5.62	58.23/44.57	27.78/ 2.14	74.52/52.48	59.70/20.74	50.04/26.89	64.17/46.28	75.67/51.23
Mean	3.00	35.64/28.15	34.49/17.35	33.35/23.48	38.23/ 8.34	34.32/18.93	33.17/26.71	40.06/32.50

Table 1. **Quantitative Results on VisA in pAP/DSC(%)**. The best results are colored red, and the second best results are underlined.

Category	AR(%)	Mem. Bank	Img Recon.	Feature Recon.				
		PatchCore [5] CVPR2022	DRAEM [9] ICCV2021	DeSTSeg [10] CVPR2023	RD4AD [1] CVPR2022	UniAD [8] NIPS2022	HVQ-Trans [4] NIPS2023	**UniAS(Ours)**
Bottle	22.82	79.17/78.42	51.32/ 1.04	72.93/65.14	68.07/47.39	69.34/67.72	72.02/65.86	84.35/79.50
Cable	14.04	51.12/51.36	9.53/ 5.35	47.50/46.35	25.68/ 5.42	48.38/27.60	50.64/31.09	78.31/73.90
Capsule	3.32	44.26/43.41	7.15/ 1.76	45.88/20.25	20.19/21.26	45.75/34.19	44.63/19.43	49.64/44.72
Hazelnut	10.06	60.12/63.23	71.82/44.48	65.71/57.62	63.56/52.25	54.35/25.84	64.21/62.40	79.17/76.80
Metal Nut	43.46	88.62/74.35	24.53/16.95	53.61/23.02	64.32/38.50	49.83/41.30	67.30/48.18	70.25/76.61
Pill	11.93	77.37/51.25	63.42/23.58	78.03/35.95	78.11/53.18	40.28/23.66	49.98/29.24	66.98/47.87
Screw	1.01	36.73/ 0.00	41.49/ 0.05	19.39/ 0.50	43.89/35.67	26.08/ 3.65	29.17/ 3.57	48.00/46.54
Toothbrush	6.40	54.03/58.46	52.60/34.09	58.06/43.04	55.51/53.96	40.09/42.81	39.80/45.56	59.41/55.42
Transistor	35.95	66.61/22.40	27.22/ 7.07	39.26/59.02	42.38/ 8.05	67.57/28.87	72.34/19.35	84.13/70.13
Zipper	7.87	53.50/53.79	73.61/57.93	61.42/49.22	57.36/63.26	33.60/21.99	37.79/32.14	57.36/55.61
Carpet	6.32	69.31/63.12	69.97/52.34	66.82/52.52	57.49/48.81	52.81/49.74	54.19/47.49	70.20/63.70
Grid	2.83	37.85/40.45	38.49/17.50	36.02/ 1.36	48.56/43.69	24.16/ 3.80	24.09/ 3.98	43.00/26.52
Leather	2.62	50.97/51.03	58.99/48.18	79.37/75.38	40.98/45.44	34.43/41.36	34.47/28.56	58.65/58.40
Tile	29.41	59.78/70.21	79.90/45.77	89.56/60.65	51.30/40.51	42.67/30.69	41.99/35.41	59.80/54.37
Wood	15.25	51.17/39.70	77.79/62.81	74.57/63.70	53.69/38.53	37.02/15.23	39.65/17.23	53.66/60.03
Mean	13.17	55.73/50.74	49.86/27.93	59.20/43.61	51.41/39.73	44.42/30.56	48.15/32.63	65.12/59.33

Table 2. **Quantitative Results on MVTec-AD in pAP/DSC(%)**. Methods are divided into Memory Bank-based, Image Reconstruction-based, and Feature Reconstruction-based categories. One-for-one methods are trained within the one-for-all scheme. Objects with logical and textural anomalies are separated. The best results are colored red, and the second best results are underlined.

B. More Experiments

B.1. Results per Class

We provide the the results of each class in our experiment on MVTec-AD (see Tab. 2) and VisA (see Tab. 1), comparing with exisiting SOTA models. We also present AR to reference the degree of imbalance in the dataset.

B.2. Ablations

We provide ablation results in pAP, DSC, and AUROC in this section, see Tab. 3 for Structural Component Study and Tab. 4 for Feature Combination Study.

We also provide additional ablation results of feature filtering. Although the effectiveness of aggregating neighbor information has been validated by previous works, we analyze the importance of our Gaussian filter in our settings,

Components								Results		
Multi-level	SAR Q.	Hybrid Structure			Filtering			Metric		
		Conv3	MG-CNN	MGG-CNN	Avg.	Gau.	Concat	pAP	DSC	AUROC
-	-	-	-	-	-	✓	✓	40.14(24.98)	25.24(34.09)	96.35(1.87)
✓	-	-	-	-	-	✓	✓	56.47(8.65)	39.06 (20.27)	96.95(1.27)
✓	✓	-	-	-	-	✓	✓	59.84(5.78)	48.35(10.98)	97.66(0.56)
✓	✓	✓	-	-	-	✓	✓	61.14(1.73)	49.27(10.06)	97.64(0.58)
✓	✓	-	✓	-	-	✓	✓	63.39(1.13)	51.56(7.77)	98.09(0.13)
✓	✓	-	-	✓	-	-	-	62.43(2.69)	52.70(6.63)	96.98(1.24)
✓	✓	-	-	✓	✓	-	-	62.79(2.33)	51.77(7.56)	98.03(0.19)
✓	✓	-	-	✓	-	✓	-	63.00(2.12)	54.26(5.07)	98.06(0.16)
✓	✓	-	-	✓	-	✓	✓	**65.12**	**59.33**	**98.22**

Table 3. **Structural Component Study** The performance gap from the default setting is shown in red.

#Levels	Levels	Metric		
		pAP	DSC	AUROC
1	{4}	44.44(20.68)	36.88(22.45)	94.83(3.39)
	{3}	53.49(11.63)	38.22(21.11)	96.37(1.85)
	{2}	56.52(8.60)	37.47(21.86)	96.16(2.06)
	{1}	40.18(24.94)	21.11(38.22)	89.30(8.92)
2	{4,3}	54.20(10.92)	45.55(13.78)	96.85(1.37)
	{4,1}	58.63(6.49)	45.63(13.70)	97.45(0.77)
3	{4,3,2}	61.57(3.55)	55.03(4.30)	97.85(0.37)
4	{4,3,2,1}	**65.12**	**59.33**	**98.22**

Table 4. **Feature Combination Study** The performance gap from the default setting is in (red). Feature levels are labeled 1,2,3,4 from lowest to highest level accordingly. $\{\cdot\}$ means the levels included.

shown in Tab. 3. Average filters (Avg. column in Tab. 3) are used in previous works [3, 5], which are proved to be too smooth to reserve necessary structural information compared to the Gaussian filter we use (Gau. column in Tab. 3) , according to the last four lines. Moreover, concatenating (Concat column in Tab. 3) the residual together with filtered features improves segmentation performance significantly, showing that incorporating details is beneficial to fine-grained localization (see last two lines).

B.3. Visualization

B.3.1 Multi-level Reconstruction

We visualize more examples of UniAS's prediction on every level in Fig. 3, labeled 4 to 1 from the highest to the lowest. These visualization result further consolidates that anomaly maps in different levels play complementary roles, detecting anomalous pixels from course to fine. Higher-level features concentrate on semantics, facilitating accurate overall anomaly localization, while lower-level features have rich textural information aiding in delineating the shape of the anomalous region.

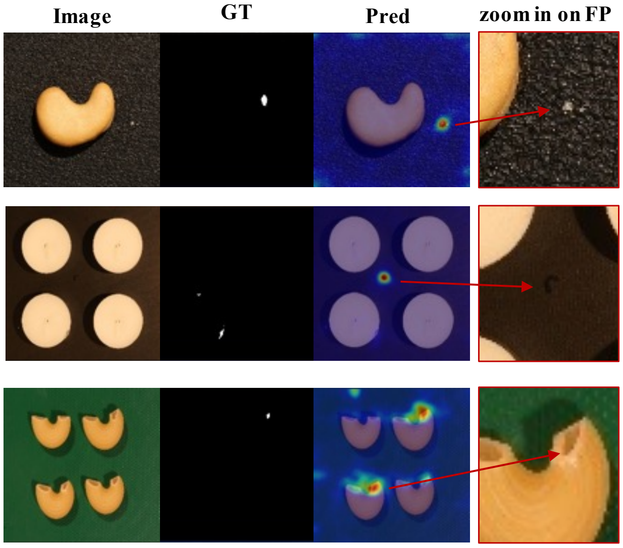

Figure 2. Examples of failed cases of our UniAS. Our model can make false predictions under noisy and complicated scenarios.

B.3.2 Failure Cases

We show some failed cases of UniAS in Fig. 2. As one can see, when the anomalous region is vague while there is obvious background noise (seen in the first two lines), our model can possibly recognize the background noise as anomaly with greater salience than the real anomaly in the foreground. Additionally, there are some noisy labels in the datatset as well (seen in the third line). This leads to false positive predictions, implying our UniAS still needs further improvement, especially under intricate and confusing situations. However, anomaly segmentation with noise is a slightly different task, which is not the topic of our work.

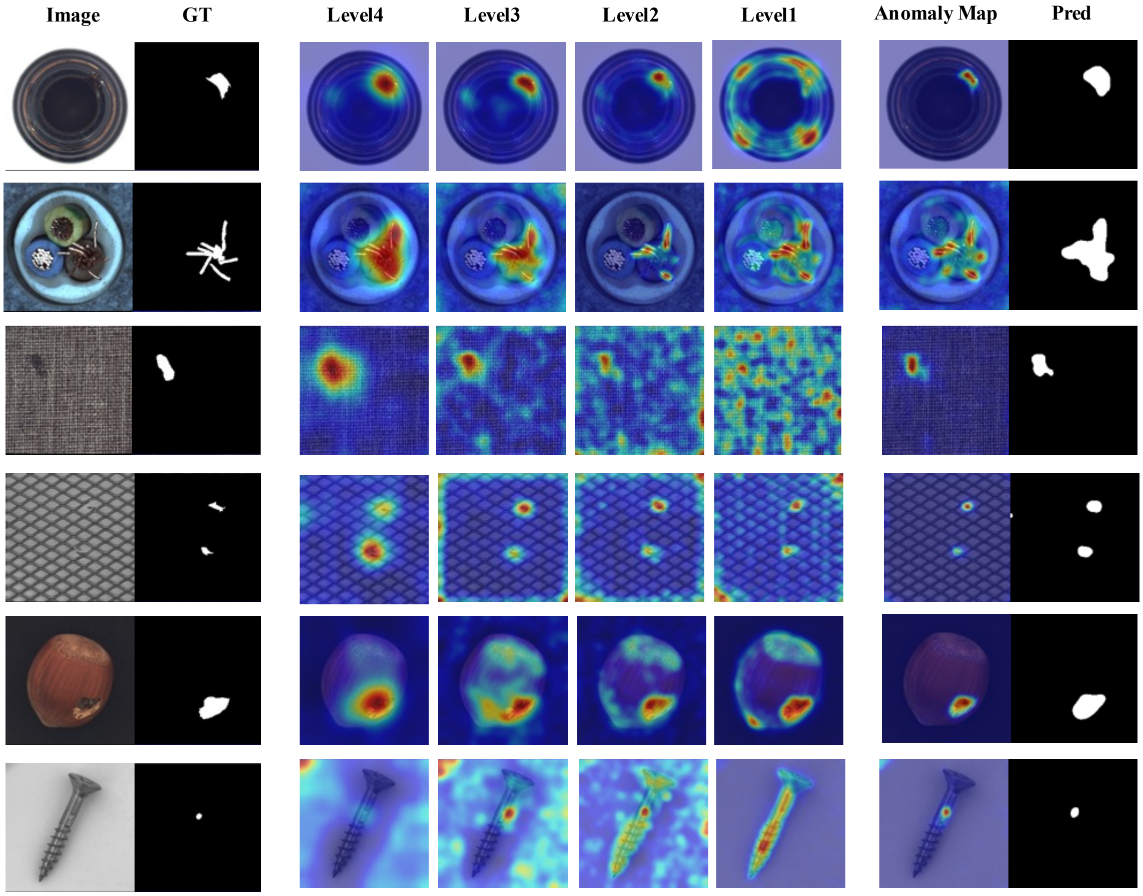

Figure 3. Additional visualization of the anomaly maps in each level and final predictions. UniAS leverages the benefits of multi-level reconstruction and achieves meaningful segmentation of anomaly.

References

[1] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 2

[2] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European Conference on Computer Vision*, pages 74–92. Springer, 2022. 1

[3] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 3

[4] Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2310.14228*, 2023. 1, 2

[5] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 2, 3

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[7] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

[8] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. 1, 2

[9] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 2

[10] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023. 2