SALAD 🥗 – Semantics-Aware Logical Anomaly Detection

Supplementary Material

This supplementary material includes additional information and visualisations. More specifically, we ablate the object composition map generation and add a further experiment to verify the importance of each branch. Ultimately, we add localisation results for MVTec LOCO and more qualitative examples.

A. Limitations and Failure Cases

Composition map creation depends on the performance of SAM-HQ and DINO, although they perform very well across diverse datasets. In the future, this can even be improved with stronger models (e.g. Perception Encoder). Additionally, there are also some cases (some are depicted in Figure 1) in which SALAD fails to detect anomalies. SALAD mostly fails on extremely near-distribution structural (Columns 1-6) and logical anomalies (Columns 7-10). Architectural improvements to the compositional and appearance branch might improve this.

B. Differences from other methods utilising composition maps

Currently, there are three different methods using composition maps – ComAD [10], CSAD [7] and PSAD [9]. SALAD's biggest difference from all three is the introduction of a specialised composition branch. This means SALAD is directly trained on the composition maps in contrast to the other three. Additionally, CSAD and PSAD require extra category-specific information, either via hand-labelled samples or via category-specific composition map procedures. SALAD performs all of this automatically without any additional information. ComAD produces composition maps of low quality, whilst SALAD produces high-quality maps.

C. Object composition map generation ablation

This section compares the design choices for the object composition map generation. First, we examine the effect of using a different feature extractor than DINO [3]. Then, we examine the importance of the number of clusters parameter.

Different feature extractor To evaluate the choice of feature extractor in component map generation, we replaced the original with other standard feature extractors: ResNet50 [6], ResNet50 DINO [3], SAM [8], ViT DINOv2 [12]. Their performance is qualitatively evaluated by comparing feature clusters C_{feat}, pseudo labels C_{pseudo}, and final composition maps C. Figure 2 depicts that ResNet50, ViT DINOv2, and ViT DINO cluster features effectively, discriminating similar objects (e.g., Columns 5 and 6). In contrast, ResNet, DINO and SAM yield poor clusters, as seen in Columns 3 and 4. This pattern continues with pseudo labels in Figure 3, where ResNet DINO and SAM exhibit loss of detail and class mismatches (Columns 7 and 8). Due to noisy pseudo labels, the lightweight semantic segmentation model struggles with generalisation (Figure 4, Columns 7 and 8). Consequently, we evaluated downstream anomaly detection performance only for ViT DINO [3] and ViT DINOv2 [12], with results detailed in the main paper.

Condition	Det. Logical	Det. Struct.	Det. Avg
Only Appearance branch	87.5 (-9.0)	94.1 (-1.6)	90.8 (-5.3)
Only Composition branch	88.1 (-8.4)	82.8 (-12.9)	85.4 (-10.7)
Only Global branch	90.8 (-5.7)	87.3 (-8.4)	89.1 (-8.1)
SALAD	96.5	95.7	96.1

Table 1. Branch importance is evaluated with the downstream performance in Anomaly detection on the MVTec LOCO dataset [2] (results are presented in AUROC). The importance is evaluated by using only one branch. The results are categorised by anomaly type, and the overall average detection rate is reported in the final column. The performance difference relative to the base model is highlighted in blue.

Different number of clusters To investigate the importance of the number of clusters during composition map generation, we qualitatively and quantitatively assessed the output composition maps. More specifically, we checked for different values of K ranging from 4 to 8. We qualitatively assessed the feature clusters, pseudo-labels and the generated composition maps. The results for different stages in the pipeline can be seen in Figure 5, Figure 6 and Figure 7. From the Figures, it can be seen that there are no significant differences, especially with the final composition maps. This would suggest that the choice of the number of clusters is robust (once it is high enough). In Figure 8, the effect of this parameter on downstream anomaly detection is depicted. All values are above the current state-of-the-art, suggesting that the parameter choice is robust.

D. Branch importance

To further show the overall importance of each branch, we evaluated the model by using one branch at a time. The results can be seen in Table 1 and in Table 2. Using only the appearance branch leads to a drop in performance of 9.0 percentage points (p. p.) for logical anomalies and 1.6 p. p. for structural anomalies. Using only the composition branch leads to a drop of 8.4 p. p. on logical anomalies and a 12.9 p. p. drop for structural anomalies. By solely using the global

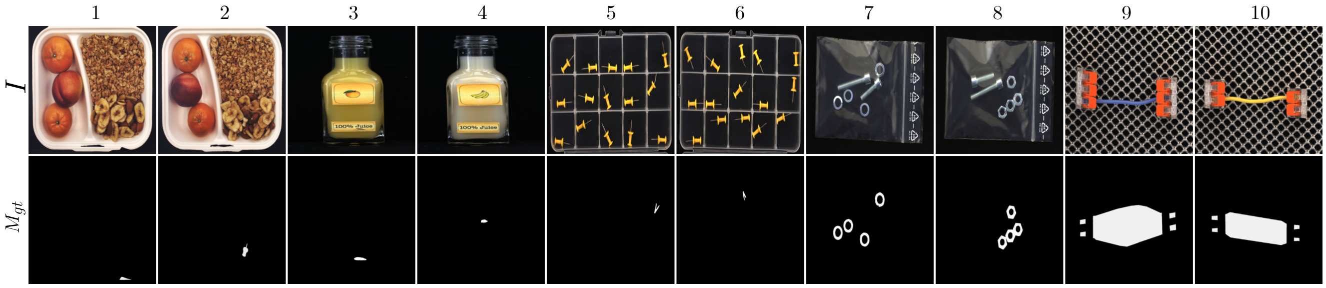

Figure 1. Failure case results. In all of the cases, SALAD produces a very low anomaly score. Most of the cases also represent near-distribution logical and structural anomalies.

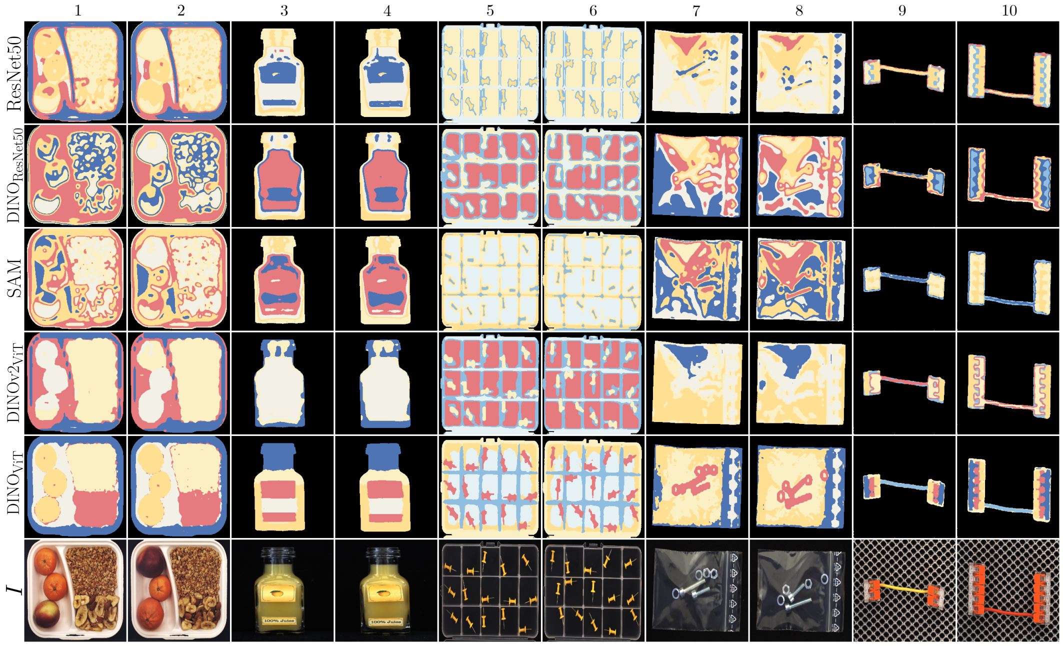

Figure 2. Qualitative comparison of the feature clusters C_{feat} produced by 5 different feature extractors: ResNet50 [6], ResNet50 DINO [3], SAM [8] and ViT DINO [3]. In the bottom row, the original image I is shown. It can be observed that both ViT DINOv2 and ViT DINO separate the objects effectively (e.g. Columns 5 and 6), while other feature extractors face problems (e.g. Columns 3 and 4).

branch, the performance drops 5.7 p. p. for logical anomalies and 8.4 p. p. for structural anomalies. The results show that the branches complement each other, especially with logical anomalies.

E. Localization results for MVTec LOCO

Following recent literature [1, 15], the AUsPRO Metric [2] is used to evaluate the localization performance. Again, it is important to highlight that most concurrent works [4, 7, 9, 10, 14] strayed away from reporting these results due to the ambiguity in pixel-level ground truths in images containing a logical anomaly. Some such cases are depicted in Figure 9. The localization results on MVTec LOCO are given in Table 3. SALAD achieves the second-highest result with an AUsPRO of 68.7%.

F. Additional qualitative results

In this section, we provide additional qualitative mask comparisons to the state-of-the-art models DRÆM [16], TransFusion [5] and EfficientAD [1]. The comparisons can be seen in Figure 10 and Figure 11. SALAD can detect more near-distribution and harder anomalies compared to previous state-of-the-art methods.

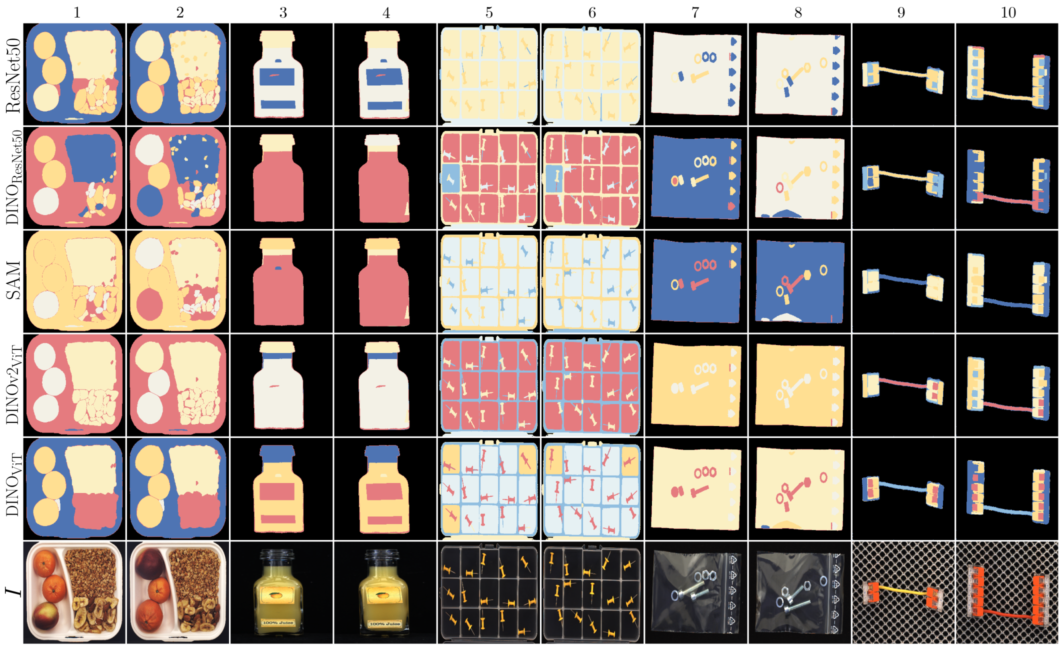

Figure 3. Qualitative comparison of the pseudo-labels C_{pseudo} produced by 5 different feature extractors: ResNet50 [6], ResNet50 DINO [3], SAM [8] and ViT DINO [3]. In the bottom row, the original image I is shown. Most methods do not face class mismatches and loss of detail except for ResNet50 DINO and SAM (e.g. Columns 7 and 8).

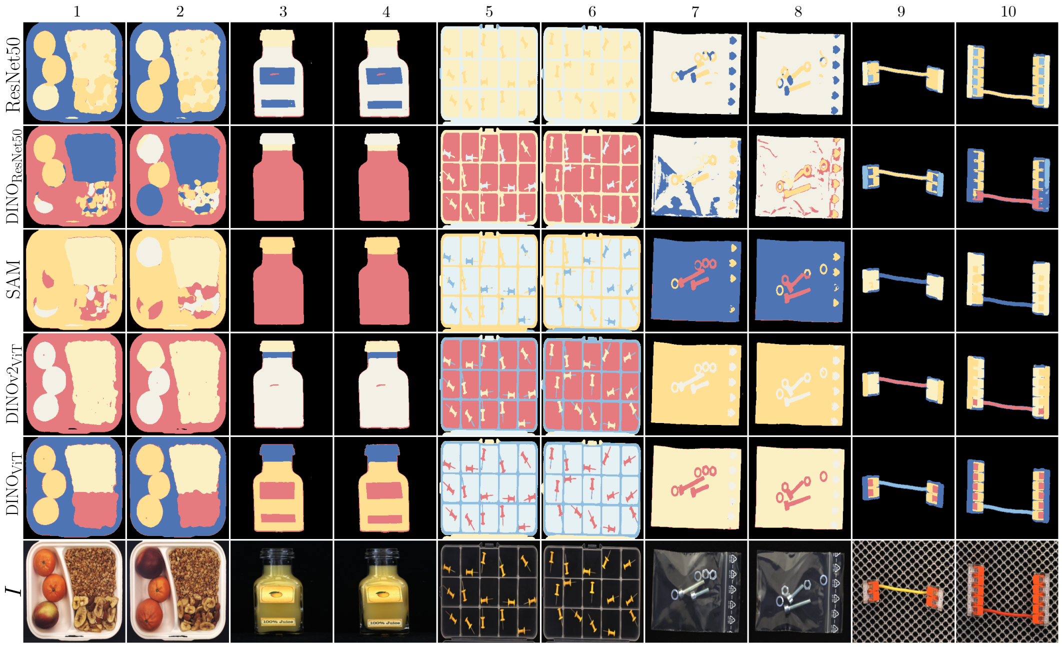

Figure 4. Qualitative comparison of the composition maps C produced by 5 different feature extractors: ResNet50 [6], ResNet50 DINO [3], SAM [8] and ViT DINO [3]. In the bottom row, the original image I is shown. While ViT DINO and ViT DINOv2 can generalise effectively, other methods face problems (e.g. Columns 7 and 8).

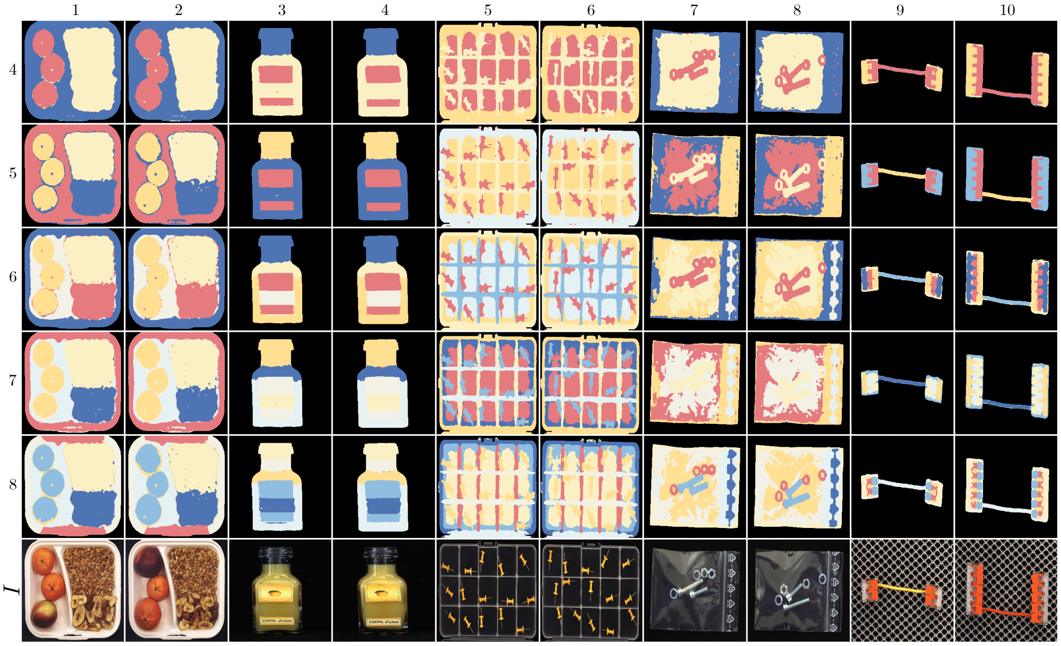

Figure 5. Qualitative comparison of the feature clusters C_{feat} produced by different numbers of clusters K (from 4 to 8). In the bottom row, the original image I is shown.

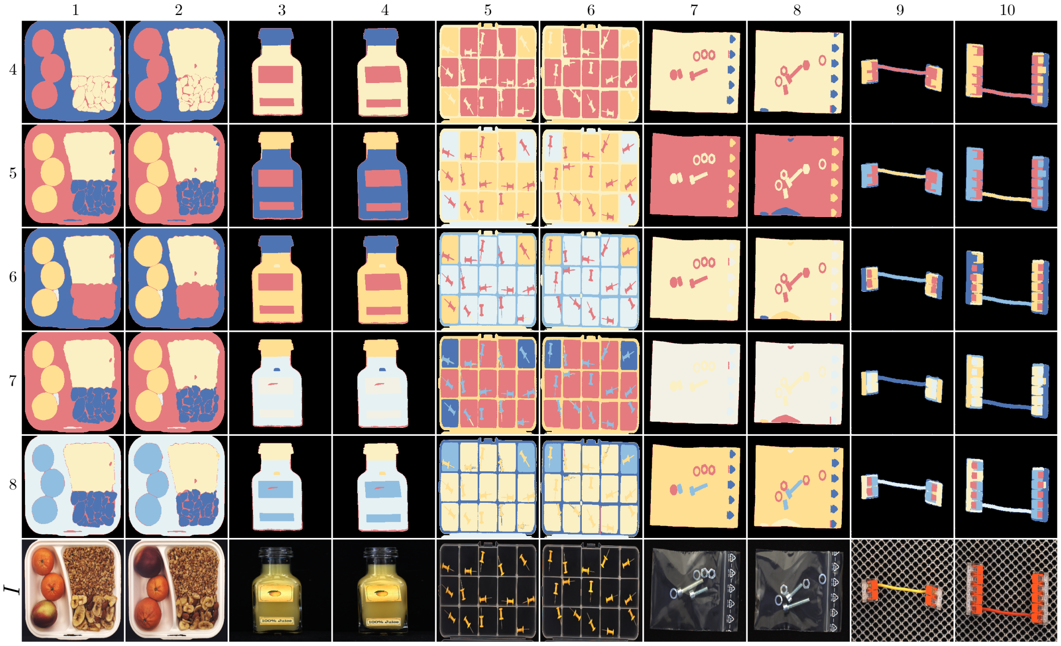

Figure 6. Qualitative comparison of the pseudo-labels C_{pseudo} produced by different numbers of clusters K (from 4 to 8).

Branch	Breakfast box	Juice bottle	Pushpins	Screw bag	Splicing conn.	*Average*
Only Appearance Branch	85.7	96.9	96.8	77.9	96.6	90.8
Only Composition Branch	77.1	87.0	87.7	88.2	86.2	85.4
Only Global Branch	82.2	97.7	91.8	86.3	87.3	89.1

Table 2. Anomaly detection (AUROC) for each branch on MVTec LOCO [2].

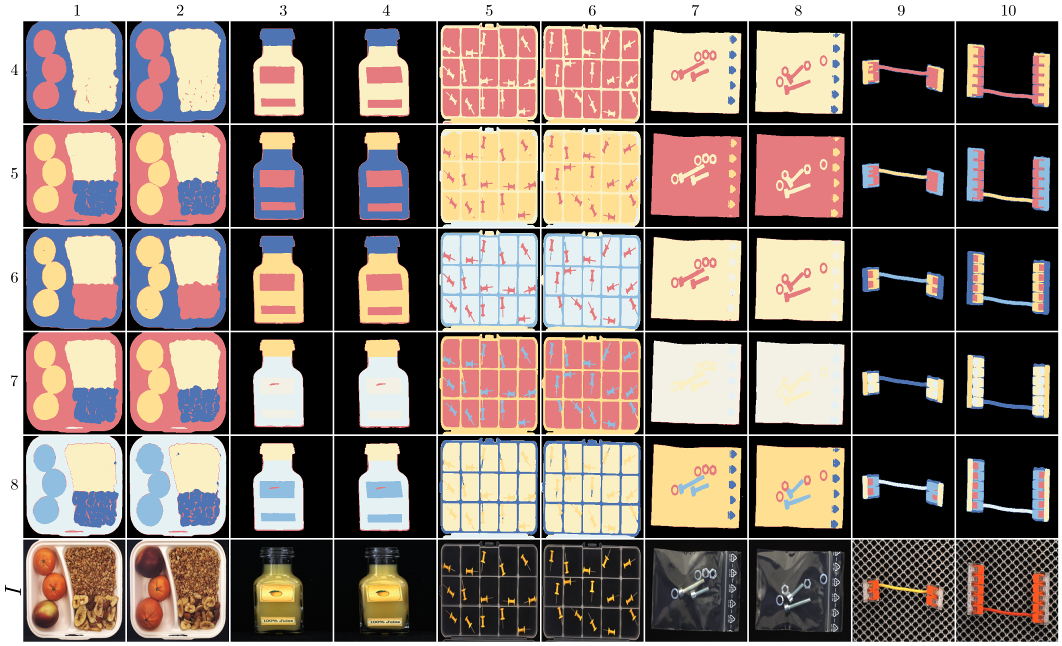

Figure 7. Qualitative comparison of the composition maps C produced by different numbers of clusters K (from 4 to 8).

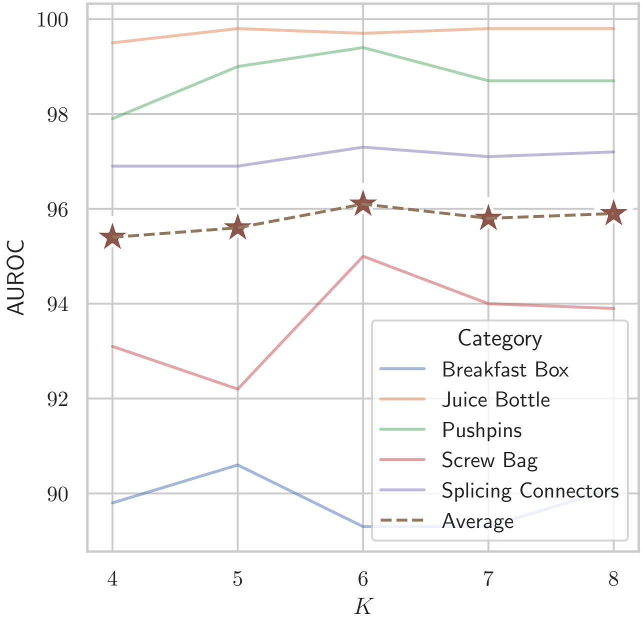

Figure 8. Anomaly detection performance on MVTec LOCO under different values for K in the object composition map generation. The default settings for K is 6.

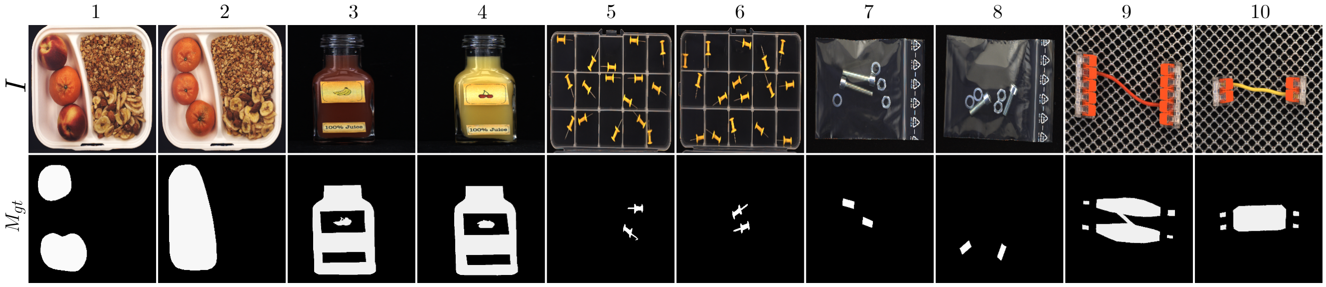

Figure 9. Examples of problematic pixel-level ground truths (M_{gt}) and their corresponding images (I) in MVTec LOCO [2] show issues with how annotations are done. The ground truths are designed to include all possible solutions, which causes ambiguity. For example, in Column 7, there are two long screws instead of one long screw and one short screw as expected. The annotation requires marking both long screws, even though marking one would still be a correct interpretation of the anomaly. This approach unfairly lowers the scores of methods that label only one screw, even if their prediction makes sense.

Category	SimpleNet [11]	DRÆM [16]	TransFusion [5]	DSR [17]	Patchcore [13]	SLSG [15]	EfficientAD [1]	*SALAD*
Breakfast box	38.8	49.9	53.5	49.9	46.6	**65.9**	60.4	49.1
Juice bottle	43.9	80.0	90.1	86.8	41.2	82.0	**93.4**	81.5
Pushpins	27.2	49.3	51.9	59.1	31.4	**74.4**	62.3	73.5
Screw bag	**66.0**	49.0	39.3	37.9	48.1	47.2	64.4	58.4
Splicing connectors	36.9	67.3	67.0	58.6	31.3	66.9	73.3	**81.2**
Average	36.3	59.1	60.4	58.5	39.7	67.3	**69.4**	68.7

Table 3. Anomaly localization (AUsPRO) on MVTec LOCO [2].

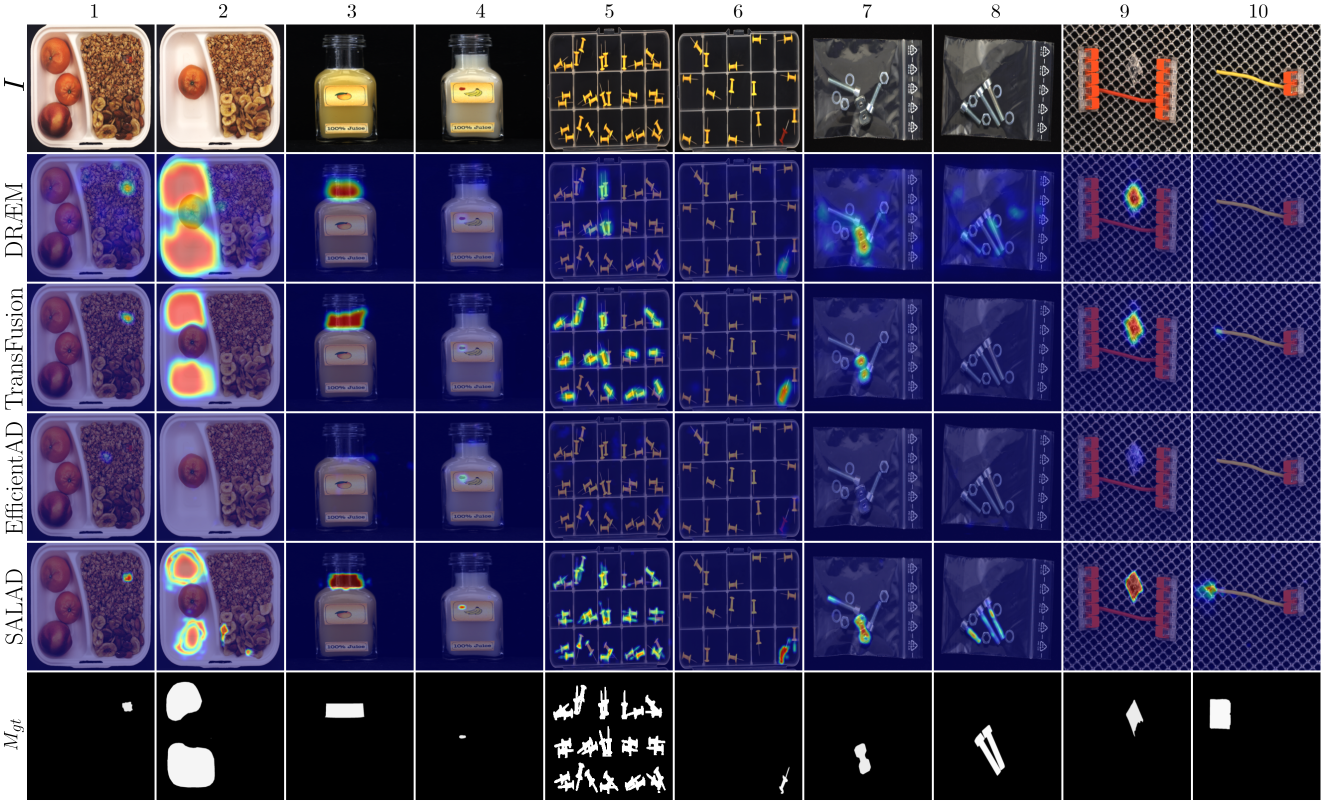

Figure 10. Qualitative comparison of the anomaly segmentation masks produced by SALAD and three other state-of-the-art methods on MVTec LOCO. In the first row, the image is shown. In the next four rows, the anomaly segmentations produced by DRÆM [16], TransFusion [5], EfficientAD [1] and SALAD are depicted, and in the last row, the ground truth mask is shown. For SALAD, we visualized the sum of A_a and A_c.

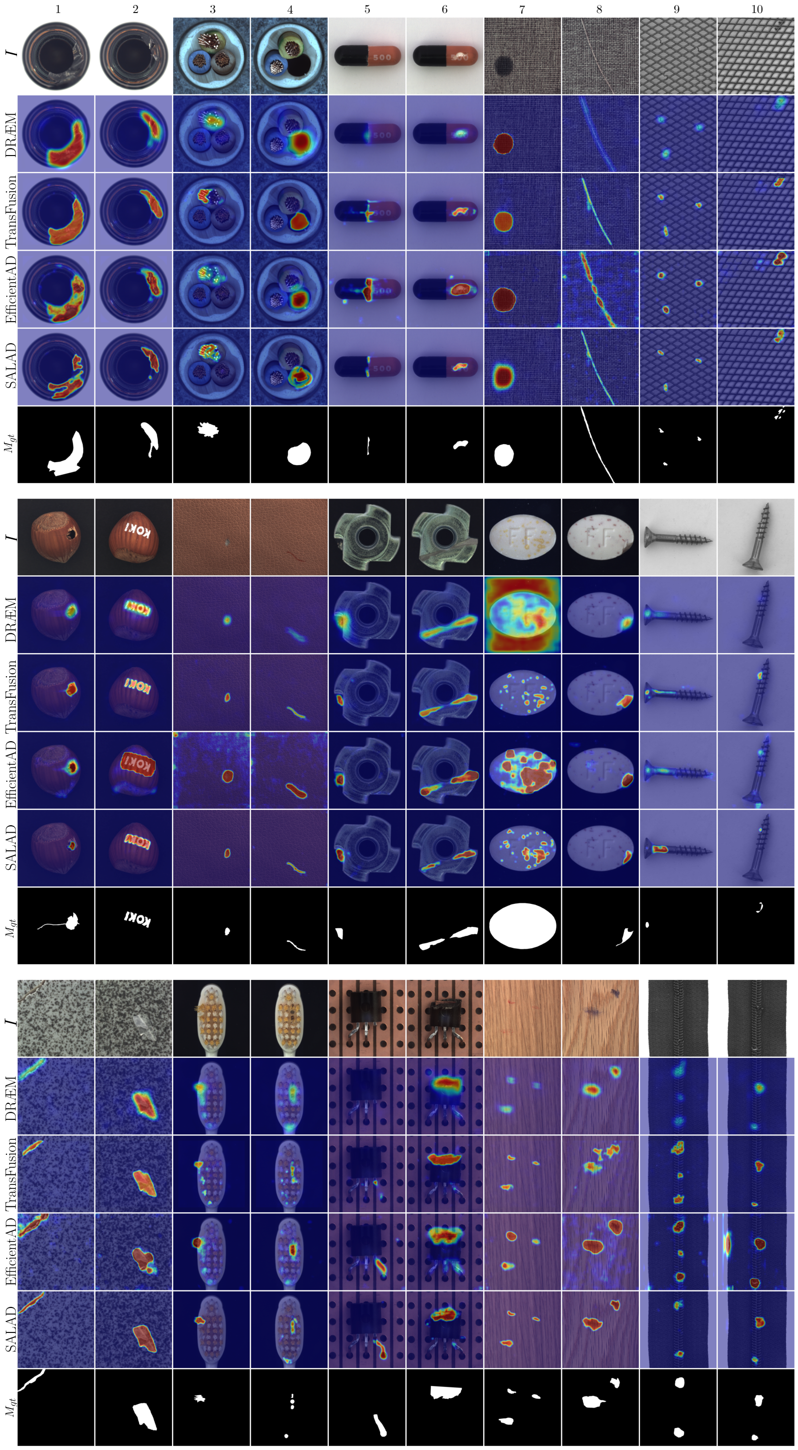

Figure 11. Qualitative comparison of the anomaly segmentation masks produced by SALAD and three other state-of-the-art methods on MVTec AD. In the first row, the image is shown. In the next four rows, the anomaly maps produced by DRÆM [16], TransFusion [5], EfficientAD [1] and SALAD are depicted, and in the last row, the ground truth mask is shown.

References

[1] Kilian Batzner, Lars Heckler, and Rebecca König. EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 128–138, 2024.

[2] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond Dents and Scratches: Logical Constraints in Unsupervised Anomaly Detection and Localization. *International Journal of Computer Vision*, 130 (4):947–969, 2022.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[4] Niv Cohen, Issar Tzachor, and Yedid Hoshen. Set Features for Anomaly Detection. *arXiv preprint arXiv:2311.14773*, 2023.

[5] Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. TransFusion – A Transparency-Based Diffusion Model for Anomaly Detection. In *European conference on computer vision*, pages 91–108. Springer, 2025.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] Yu-Hsuan Hsieh and Shang-Hong Lai. CSAD: Unsupervised component segmentation for logical anomaly detection. In *In Proceedings of the British Machine Vision Conference (BMVC)*, 2024.

[8] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Soopil Kim, Sion An, Philip Chikontwe, Myeongkyun Kang, Ehsan Adeli, Kilian M Pohl, and Sang Hyun Park. Few shot part segmentation reveals compositional logic for industrial anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8591–8599, 2024.

[10] Tongkun Liu, Bing Li, Xiao Du, Bingke Jiang, Xiao Jin, Liuyi Jin, and Zhuo Zhao. Component-aware anomaly detection framework for adjustable and logical industrial visual inspection. *Advanced Engineering Informatics*, 58:102161, 2023.

[11] 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.

[12] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[13] 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.

[14] Shota Sugawara and Ryuji Imamura. PUAD: Frustratingly simple method for robust anomaly detection. In *2024 IEEE international conference on image processing (ICIP)*, 2024.

[15] Minghui Yang, Jing Liu, Zhiwei Yang, and Zhaoyang Wu. SLSG: Industrial Image Anomaly Detection by Learning Better Feature Embeddings and One-Class Classification. *Pattern Recognition*, 156:110862, 2024.

[16] 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.

[17] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DSR–A dual subspace re-projection network for surface anomaly detection. In *European Conference on Computer Vision*, pages 539–554. Springer, 2022.