SuperAD: A Training-free Anomaly Classification and Segmentation Method for CVPR 2025 VAND 3.0 Workshop Challenge Track 1: Adapt & Detect

Huaiyuan Zhang[1] Hang Chen[1] Yu Cheng[1] Shunyi Wu[1] Linghao Sun[1]
Linao Han[1] Zeyu Shi[2] Lei Qi[1*]
[1]School of Computer Science and Engineering, Southeast University, China
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
{zhang_hy, hangchen, chengyu, shunyiwu, linghaosun, linaohan, qilei}@seu.edu.cn
shizeyu@njust.edu.cn

Abstract

*In this technical report, we present our solution to the CVPR 2025 Visual Anomaly and Novelty Detection (VAND) 3.0 Workshop Challenge Track 1: Adapt & Detect: Robust Anomaly Detection in Real-World Applications. In real-world industrial anomaly detection, it is crucial to accurately identify anomalies with physical complexity, such as transparent or reflective surfaces, occlusions, and low-contrast contaminations. The recently proposed MVTec AD 2 dataset significantly narrows the gap between publicly available benchmarks and anomalies found in real-world industrial environments. To address the challenges posed by this dataset—such as complex and varying lighting conditions and real anomalies with large scale differences—we propose a fully training-free anomaly detection and segmentation method based on feature extraction using the DINOv2 model named **SuperAD**. Our method carefully selects a small number of normal reference images and constructs a memory bank by leveraging the strong representational power of DINOv2. Anomalies are then segmented by performing nearest neighbor matching between test image features and the memory bank. Our method achieves competitive results on both test sets of the MVTec AD 2 dataset.*

1. Introduction

1.1. Background

Unsupervised anomaly detection and localization has emerged as a key technology in computer vision, with wide-ranging applications in real-world scenarios such as industrial quality inspection and autonomous driving. The central challenge of this task lies in training models solely on normal samples, while enabling accurate identification and precise localization of previously unseen defects during testing. In recent years, fueled by the advancement of deep learning, numerous methods have achieved remarkable progress on mainstream benchmark datasets such as MVTec AD and VisA. However, as model performance on these datasets approaches saturation—for instance, segmentation AU-PRO scores of some algorithms on MVTec AD have surpassed 97%—their limitations have become increasingly apparent:

1. **Limited Scene Diversity**: Existing datasets primarily focus on objects with clear textures and simple structures, lacking coverage of complex industrial scenarios involving transparent or reflective surfaces (e.g., glassware, metal products), as well as bulk, overlapping items (e.g., granular materials).
2. **Idealized Defect Types**: Most defects in current datasets are large, centrally located anomalies, overlooking real-world industrial defects such as edge anomalies, subpixel-level scratches (e.g., hairline cracks), and low-contrast contaminations (e.g., transparent foreign objects).
3. **Insufficient Environmental Robustness**: These datasets often ignore variations in lighting conditions (e.g., dark-field, backlighting, overexposure), resulting in models with limited generalization ability when deployed across different devices or under varying environments.

1.2. Challenge Description

As a next-generation benchmark for industrial anomaly detection, the MVTec AD 2 dataset [6] systematically addresses the aforementioned limitations by introducing eight complex and diverse scenarios. Its core features include:

- **Simulation of Physical Complexity**:
  - *Transparent and reflective surfaces*: Categories such

*Corresponding Author

as `Vial` feature liquid refraction artifacts, and `Sheet Metal` includes mirror-like reflections, both of which challenge the model's ability to reason about light propagation.
  - *Bulk and overlapping objects*: Examples like `Wallplugs` and `Walnuts` involve random occlusions and truncated boundaries between objects, requiring semantic-level understanding.
  - *High intra-class variability in normal samples*: Categories such as `Fabric` (with diverse textures) and `Can` (with geometric pattern deformations) demand models to learn tight boundaries of the normal data manifold.
- **Verification of Detection Limits**:
  - *Tiny objects and boundary anomalies*: Plastic contaminants occupying less than 0.1% of the image area in the `Rice` category, and missing regions at the image boundaries in `Wallplugs` pose significant challenges to the model's perceptual capability at high resolution.
  - *Implicit structural consistency*: `Fruit Jelly` require the model to assess the plausibility of ingredient distributions, despite the absence of explicit logical constraints.
- **Cross-Domain Generalization Evaluation**: Each scene includes at least four lighting conditions (regular, underexposed, overexposed, and additional light sources), simulating distribution shifts caused by device variations in real-world production environments. This enables systematic evaluation of model robustness.

On the MVTec AD 2 dataset, mainstream methods exhibit clear performance bottlenecks:

Limited localization capability: EfficientAD [1] and PatchCore [10] achieve average AU-PRO$_{0.05}$ scores [6] of only 58.7% and 53.8%, respectively. In highly complex scenarios such as `Can` and `Rice`, their performance drops below 30%.

Poor robustness: MSFlow [14] shows a significant performance degradation of 51.1% in AU-PRO$_{0.05}$ on the mixed lighting test set $TEST_{priv,mix}$ compared to the standard test set $TEST_{\text{priv}}$, highlighting its sensitivity to environmental variations.

In summary, this paper targets the key challenges highlighted in the MVTec AD 2 dataset and aims to:

- Develop a novel and highly robust model capable of accurately detecting subtle defects within highly variant normal samples;
- Enhance localization performance on small targets and structurally complex objects, such as transparent or overlapping instances;
- Strengthen the model's generalization ability under varying illumination conditions, thereby improving its practical applicability in industrial deployment.

2. Methodology

2.1. Model Design

2.1.1. Approach

In recent years, memory bank-based methods have demonstrated remarkable performance on various anomaly detection and segmentation benchmarks. PatchCore [10] introduces a greedy coreset selection mechanism to construct a compact memory bank, significantly reducing the number of features while preserving the overall distribution of the original feature space. This design effectively balances detection accuracy and retrieval efficiency. DMAD [7] unifies two commonly encountered scenarios in industrial settings: one with only normal samples available and the other with a limited number of labeled anomalies. It achieves this by constructing a normal memory bank and an expandable anomaly memory bank, which store features of normal and observed anomalous patterns, respectively, thereby improving adaptability to real-world complexities.

It is worth noting that the success of these memory-based approaches heavily depends on the quality of feature extraction. As a result, most methods rely on powerful pre-trained visual backbones such as WideResNet [12] or Vision Transformer (ViT) [5]. Prior studies reveal that different layers in deep models capture distinct types of information: shallow layers tend to focus on local high-frequency details (e.g., textures, edges), whereas deeper layers encode more abstract semantic information. Therefore, combining both shallow and deep features enables the model to capture both global structure and local details, which is crucial for accurate anomaly detection.

For instance, PaDiM [3] employs a pre-trained ResNet to extract patch-wise features from four different layers and models the distribution of normal features at each spatial location using a multivariate Gaussian distribution, parameterized by the mean and covariance. APRIL-GAN [2], on the other hand, leverages CLIP's powerful multimodal alignment capabilities. It extracts four-layer features from both the test image and its corresponding normal reference image, performs layer-wise matching, and averages the results to localize anomalous regions.

Fortunately, recent advances in self-supervised learning, such as DINOv2 [9], have shown strong capability in capturing rich semantic information for visual tasks. Building upon this, our method constructs a class-specific normal feature memory bank for each category in the MVTec AD 2 dataset. During inference, the features from various regions of an input image are matched against those stored in the memory bank to detect anomalies. We adopt the powerful DINOv2 backbone to extract multi-level features, aiming to achieve accurate and fine-grained anomaly segmentation.

8个场景、每个场景选16个样本

用 DINOv2-large 分别提取 “测试图像” 和 “正常参考图像” 的多层特征

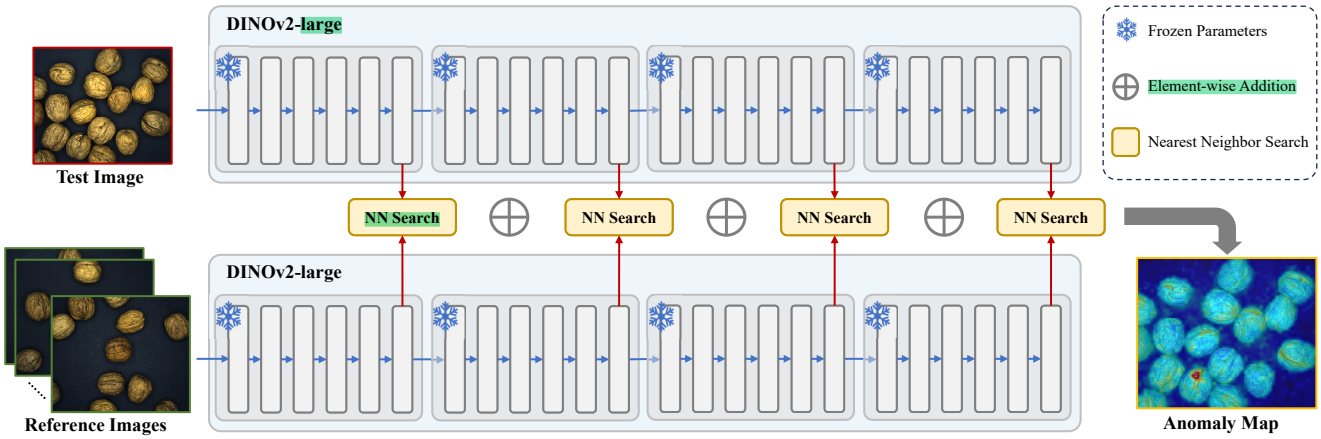

Figure 1. The overall architecture of our proposed method.

2.1.2. Architecture

The overall architecture of our proposed method is illustrated in Figure 1.

For each category in the dataset, we construct a memory bank consisting of 16 normal reference images. The selection of these reference images follows a two-step procedure. First, we extract CLS token representations from all training images using the DINOv2 model. Then, the same greedy coreset selection method as used in Patchcore is applied to group these feature vectors into 16 clusters. This strategy maximizes the coverage of diverse normal patterns within a category without changing the distribution of all features, thereby enhancing the representativeness of the reference set and reducing false positive rates.

For each test image, we first extract multi-level features using the DINOv2 model. For each level of features, we compute the similarity between the test image features and those stored in the memory bank to retrieve the nearest neighbor information. Our method is based on the assumption that normal regions in the test image tend to find similar regions among the reference images, while anomalous regions lack such matches. By evaluating the similarity at each spatial location, we obtain an anomaly map for each level. Finally, these anomaly maps are averaged and upsampled to the original resolution to generate the final anomaly segmentation map.

2.1.3. Training

Our proposed method requires no training. For each category in the MVTec AD 2 dataset, we construct a few-shot feature memory bank of normal reference images using samples from the training set. The number of reference samples is fixed at 16. We adopt the pretrained DINOv2-ViT-L-14 model as the feature extractor, which consists of 24 transformer layers and approximately 300 million parameters. Features are extracted from four specific layers (i.e., layer 6, 12, 18, and 24) to generate the final anomaly segmentation map.

To ensure memory efficiency, input image resolutions are adjusted: for all categories except `Sheet Metal`, the shorter side is resized to 672 pixels while preserving the original aspect ratio. For `Sheet Metal`, due to the elongated image size, the shorter side is resized to 448 pixels, also preserving the original aspect ratio. This configuration allows our method to run entirely on a single 24GB GPU (e.g., NVIDIA GeForce RTX 3090).

For the `Vial` and `Wallplugs` categories, we further enhance segmentation performance by extracting foreground features from the input images, as detailed in Section 2.2.1.

For the `Fabric` and `Walnuts` categories, there are cases where the central pattern is entirely normal while anomalies exist only at the edges (e.g., in the `Fabric` category, a piece of fabric may be placed over the background fabric). After the initial detection, we apply post-processing to these two categories by filling the interiors of closed regions, further improving prediction accuracy.

2.2. Dataset & Evaluation

2.2.1. Dataset Utilization

Since our proposed method is training-free, for each category, we only utilize images from its training set to construct a few-shot memory bank. In this method, we innovatively propose an adaptive background mask generation technique based on Principal Component Analysis (PCA) and morphological optimization for the image preprocessing stage. This enhances visual feature representation while suppressing redundant background information.

First, we apply PCA to reduce the dimensionality of features extracted from the input image and extract the first principal component. This technique captures the direction of maximum variance in the feature space by perform-

ing singular value decomposition on the covariance matrix, which is mathematically expressed as:

$$\mathbf{PC}_1 = \arg\max_{\|\mathbf{v}\|=1} \mathrm{Var}(\mathbf{X}\mathbf{v}), \tag{1}$$

where $\mathbf{X}$ denotes the normalized feature matrix, and $\mathbf{v}$ is the projection vector. The first principal component reflects the primary mode of variation in the feature distribution. Subsequently, we binarize the projection values based on a predefined threshold τ to generate an initial mask:

$$\mathcal{M}_{\text{init}}[i] = \begin{cases} 1 & \text{if } \mathbf{PC}_1[i] > \tau \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

Since it is initially unclear which part corresponds to foreground or background, we propose an adaptive decision strategy based on variance analysis. Specifically, we initially designate one region as foreground and the other as background, then compute the feature variance within each region and compare their medians. If the variance of the foreground region is lower than that of the background region, it suggests a potential misclassification, and we invert the mask accordingly. This strategy effectively mitigates foreground-background misclassification caused by improper thresholding.

$$\mathcal{M}_{\text{init}} = \begin{cases} \mathcal{M}_{\text{init}} & \text{if } \mathrm{MD}(\mathrm{Var}(\mathbf{F}_{\text{msk}})) \geq \mathrm{MD}(\mathrm{Var}(\mathbf{F}_{\text{n-msk}})) \\ \neg\mathcal{M}_{\text{init}} & \text{otherwise} \end{cases}, \tag{3}$$

where M_{init} denote the initial binary mask (1 for masked region, 0 for unmasked region), $F_{\text{msk}} \in \mathbb{R}^{N_{\text{msk}} \times D}$ and $F_{\text{n-msk}} \in \mathbb{R}^{N_{\text{n-msk}} \times D}$ represent the feature matrices of the masked and unmasked regions, respectively. $\mathrm{Var}(\cdot)$ computes the per-channel variance (dimension: $D \times 1$), and $\mathrm{MD}(\cdot)$ returns the median value of the variance vector to suppress the influence of outliers. The operator $\neg$ denotes boolean mask inversion.

To address discrete noise in the initial mask, we apply 2D morphological operations for post-processing. Specifically, we use a $k \times k$ square kernel to dilate the mask, enhancing region connectivity. By combining dilation and erosion, we eliminate holes and smooth the boundaries. This process is formally defined as:

$$\mathcal{M}_{\text{final}} = \mathrm{Closing}\left(\mathrm{Dil.}\left(\mathcal{M}_{\text{init}} \in \{0,1\}^{H \times W}\right), \mathcal{K}\right), \tag{4}$$

where $\mathcal{K}$ is the morphological kernel and $H \times W$ is the spatial dimension of the feature grid. Finally, the optimized 2D mask is reshaped back to the feature vector dimension, yielding a boolean mask matrix used for element-wise filtering of the original features. This effectively suppresses interference from low-information background regions.

Object	AU-ROC$_{0.05}$	F1 Score
Can	58.61	0.18
Fabric	68.29	28.22
Fruit Jelly	80.93	48.27
Rice	92.92	68.44
Vial	69.04	35.88
Wallplugs	77.90	19.19
Walnuts	89.23	75.05
Sheet Metal	76.80	40.13
Mean	76.71	39.42

Table 1. AU-ROC$_{0.05}$ and segmentation F1 score (in%) on binarized images for *TEST$_{public}$* set.

In all experiments, we set the number of PCA components to 1, the threshold τ to 1.0, and the morphological kernel size to 3×3. These parameters generalize well across various datasets and require no dataset-specific tuning. Experimental results demonstrate that the proposed preprocessing method preserves the integrity of the main object features while significantly reducing background noise interference, thereby improving model performance.

2.2.2. Evaluation Criteria

We evaluate the model's performance primarily using the **pixel-level F1 score**. This metric combines precision and recall by computing their harmonic mean, thereby providing a balanced measure of the model's ability to detect anomalies at the pixel level. Specifically, the F1 score is calculated as:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{5}$$

where Precision denotes the proportion of predicted anomalous pixels that are truly anomalous, while Recall represents the proportion of actual anomalous pixels that are correctly identified by the model.

During evaluation, we optimize the F1 score by adjusting the decision threshold on *TEST$_{public}$* dataset to determine the optimal boundary for anomaly segmentation. This process ensures that the model achieves a balanced trade-off between precision and recall.

The use of the pixel-level F1 score as the evaluation metric enables precise assessment of the model's capability to identify anomalous regions in complex image data. Notably, the MVTec AD 2 dataset emphasizes the detection of **small defects**, which may occupy only a few pixels in an image. In such cases, conventional metrics like the Area Under the Receiver Operating Characteristic Curve (AU-ROC) can be dominated by larger defects, thus failing to accurately reflect the model's performance on smaller anomalies. In contrast, the pixel-level F1 score places equal

Object	PatchCore [10]	RD [4]	RD+ [11]	EfficientAD [1]	MSFlow [14]	SimpleNet [8]	DSR [13]	Ours
Can	0.3 / 0.1	0.1 / 0.1	0.1 / 0.1	0.8 / 0.1	5.0 / 0.1	0.6 / 0.1	0.4 / 0.1	**17.3 / 1.9**
Fabric	11.5 / 9.8	2.6 / 2.2	2.9 / 2.3	7.6 / 1.0	22.0 / 4.1	21.6 / 10.2	7.9 / 5.0	**77.4 / 65.3**
Fruit Jelly	8.7 / 8.2	22.5 / 22.7	26.9 / 26.7	20.8 / 18.2	**47.6** / 38.1	25.1 / 23.0	17.9 / 17.2	41.3 / **40.9**
Rice	3.8 / 4.2	7.0 / 3.9	9.5 / 2.9	15.0 / 0.5	19.1 / 1.8	11.6 / 1.0	1.5 / 1.4	**60.9 / 61.2**
Sheet Metal	1.8 / 1.1	41.3 / 39.2	40.9 / 37.7	9.3 / 3.8	13.0 / 7.6	14.6 / 2.8	13.9 / 14.4	**59.5 / 59.7**
Vial	2.3 / 2.2	28.0 / 28.3	28.2 / 22.8	30.5 / 26.5	23.3 / 6.2	31.9 / 17.5	28.2 / 27.9	**42.8 / 40.8**
Wallplugs	0.0 / 0.0	1.9 / 0.8	1.3 / 0.9	4.4 / 0.3	0.1 / 0.2	1.0 / 0.3	0.4 / 0.4	**13.7 / 6.7**
Walnuts	1.2 / 1.3	41.2 / 36.7	44.1 / 40.5	34.6 / 13.3	44.5 / 14.3	35.2 / 14.3	17.0 / 9.6	**69.1 / 69.1**
Mean	3.7 / 3.4	18.1 / 16.7	19.2 / 16.7	15.4 / 8.0	21.8 / 9.0	17.7 / 8.7	10.9 / 9.5	**47.8 / 43.2**

Table 2. Performance comparison of segmentation F1 score (in%) on binarized images for $TEST_{priv}$ / $TEST_{priv,mix}$ set. The best results are highlighted in bold.

emphasis on detecting all anomalous regions, regardless of their size. This makes it particularly well-suited to the challenges posed by the MVTec AD 2 dataset, as it can reliably evaluate the model's ability to correctly identify even the smallest defects.

3. Results

3.1. Performance Metrics

Our proposed method demonstrates excellent performance in pixel-level anomaly detection tasks. Specifically, on the $TEST_{public}$ dataset, the model's performance in terms of AUROC$_{0.05}$ and segmentation F1 score is summarized in Table 1.

Moreover, based on the official test results from the VAND 3.0 Challenge server, our model achieves an F1 score of **47.18%** on the $TEST_{priv}$ dataset and **42.51%** on the $TEST_{priv,mix}$ dataset. These results indicate that the model achieves a good balance between precision and recall, and can effectively detect anomalous pixels in images. Notably, the model maintains relatively stable performance even under distributional shifts in the data.

To more comprehensively evaluate the model's performance, we also consider other key metrics. On the $TEST_{priv}$ dataset, the model achieves an AucPro$_{0.05}$ score of **60.51%**, while on the $TEST_{priv,mix}$ dataset, the score is **58.37%**. These results suggest that the model has strong detection capabilities for anomalous regions across different thresholds and remains accurate even under variations in illumination and other environmental factors.

In addition, for image-level classification tasks, the model achieved ClassF1 scores of **70.2%** and **74.4%** on the $TEST_{priv}$ and $TEST_{priv,mix}$ datasets, respectively, demonstrating strong capability in distinguishing between normal and anomalous samples at the image level.

3.2. Comparison

Due to the recent release of the MVTec AD 2 dataset and the unavailability of Ground Truth for the $TEST_{priv}$ and $TEST_{priv,mix}$ test sets, we compare our proposed method with other approaches listed in the MVTec AD 2 dataset paper, as shown in Table 2. On both test sets, our method, SuperAD, consistently outperforms previous methods. Moreover, our method requires no training, demonstrating superior generalization capabilities compared to the other approaches.

4. Discussion

4.1. Challenges & Solutions

During our experiments, we observe that certain categories (such as `Fruit Jelly`, `Vial` and `Wallplugs`) exhibit high intra-class variability. When constructing the reference feature memory bank using randomly selected images for these categories, many normal regions are incorrectly identified as anomalies due to the lack of sufficiently similar patterns in the memory bank. To address this issue, we propose selecting reference images using a greedy coreset selection strategy rather than random sampling. This approach increases the diversity of patterns within the memory bank and helps reduce false positives. A detailed explanation of this method is provided in Section 2.1.2.

We further analyze the anomaly segmentation maps generated from the similarity scores between the extracted features at four different layers and the reference images for each category. We observe that, due to the relatively clear separation between foreground and background in most categories of the MVTec AD 2 dataset, false positives in background regions are generally limited. However, in the case of the `Wallplugs` category, some false detections still occur in the background owing to its highly complex and diverse patterns.

To gain deeper insights, we evaluate the effectiveness of PCA-based binary classification on the features extracted by DINOv2. Our results indicate that applying PCA to the shallow layers of DINOv2, with a threshold set to 1, effectively separates foreground and background regions. More details can be found in Section 2.2.1.

Based on this observation, we apply a foreground feature

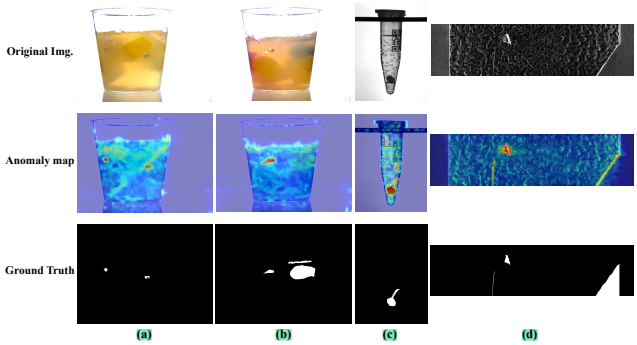

Figure 2. Examples of typical failure cases: (a) false positives on air bubbles; (b) missed detection of blurred objects; (c) false positives due to specular highlights; (d) missed detections of missing-type anomalies.

extraction preprocessing strategy to the `Wallplugs` category, which leads to improved segmentation performance for this particularly challenging class.

4.2. Model Robustness & Adaptability

It is worth emphasizing that our proposed method is entirely training-free, and thus does not require fine-tuning DINOv2 for any specific category. As a result, the generalization ability of the model remains fully preserved.

Our approach relies heavily on the powerful feature extraction capabilities of DINOv2, which enables the extraction of semantically rich representations suitable for comparison across diverse categories.

During the construction of the memory bank, we not only select 16 reference images but also employ a greedy coreset selection strategy to ensure that the selected images are as diverse as possible. This strategy effectively alleviates the issue of limited pattern diversity within the memory bank, which might otherwise fail to represent the full range of intra-class variability. Consequently, this design further enhances the robustness of our model.

4.3. Future Work

Although our method achieves competitive performance, several prominent issues remain unresolved and warrant further investigation.

False positives on air bubbles: As illustrated in Figure 2(a), `Fruit Jelly` category contains a wide variety of air bubbles with high uncertainty. These false positives are primarily caused by the diversity of bubble appearances, which often do not match the bubble patterns in the fixed memory bank of normal features. In the future, the robustness of the model to complex pattern variations could be enhanced to reduce such misclassifications.

Missed detections of reflections and blurry objects: As shown in Figure 2(b), in `Fruit Jelly` category, light-colored reflections and light scattering caused by the optical properties of jelly can diminish the perceived abnormality of dark objects. The model may erroneously match such regions to normal patterns. Future work could improve the model's ability to recognize objects within transparent or semi-transparent media to reduce missed detections.

False positives on specular highlights: As depicted in Figure 2(c), in `Vial` and `Can` categories, specular highlights caused by illumination often lead to false predictions. Incorporating advanced illumination compensation or highlight removal techniques may enhance the model's robustness to such lighting artifacts and reduce false positives.

Missed detections of missing-type anomalies: As shown in Figure 2(d), in `Sheet Metal` and `Vial` categories, missing-type anomalies are sometimes difficult to detect due to their high visual similarity with the background. Future research may explore ways to enable model to better capture object completeness features, thereby improving its ability to detect missing-type anomalies.

5. Conclusion

In this report, we propose a fully training-free anomaly detection and segmentation method that achieves robust performance under complex and varying lighting conditions. In our method, we first employ a greedy coreset selection strategy to select a small number of diverse normal reference images. Then, leveraging the powerful representational capacity of the DINOv2 model, we extract image features from the selected references to construct a memory bank. For a test image, we extract multi-scale features and perform nearest neighbor matching with the memory bank at each scale to generate anomaly segmentation maps. These maps are then averaged to produce the final result.

Our method demonstrates that the large pre-trained model DINOv2 possesses excellent image representation capabilities. Relying solely on these capabilities without any fine-tuning enables effective performance on dense prediction tasks such as anomaly segmentation.

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, 5

[2] Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023. 2

[3] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Inter-*

national conference on pattern recognition, pages 475–489. Springer, 2021. 2

[4] 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. 5

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[6] Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. *arXiv preprint arXiv:2503.21622*, 2025. 1, 2

[7] Jianlong Hu, Xu Chen, Zhenye Gan, Jinlong Peng, Shengchuan Zhang, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Liujuan Cao, and Rongrong Ji. Dmad: Dual memory bank for real-world anomaly detection. *arXiv preprint arXiv:2403.12362*, 2024. 2

[8] 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. 5

[9] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[10] 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, 5

[11] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24511–24520, 2023. 5

[12] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 2

[13] 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. 5

[14] Yixuan Zhou, Xing Xu, Jingkuan Song, Fumin Shen, and Heng Tao Shen. Msflow: Multiscale flow-based framework for unsupervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2, 5