ISVL++: Towards a Unified Framework for Accurate Anomaly Localization in Challenging Industrial Settings

Xing-ao Wang[1] * Zijun Zhou[1] Yibo Zhang[1] Zihe Xiao[1] Zifan Shen[2]
Cheng Tao[3] Xiaonan Wang[3] Tiancan Mei[4]
[1]School of Remote Sensing and Information Engineering, Wuhan University
[2]School of Artificial Intelligence, Wuhan University
[3]Hexagon Manufacturing Intelligence
[4]School of Electronic Information, Wuhan University
[1,2,4]{xingaowang, zzjivan, zybWHU, xiaozihe, aiziqiangSZF, mtc}@whu.edu.cn
[3]{cheng.tao, xiaonan.x.wang}@hexagon.com

Submission Information: Industrial Track, Regular setting

Abstract

In this project, we present a unified anomaly detection framework for the CVPR Workshop 2026 Visual Anomaly and Novelty Detection Challenge, with a focus on unsupervised defect detection in complex industrial environments. Our method is built upon a reconstruction-based anomaly detection paradigm and further incorporates anomaly synthesis to enhance the model's ability to distinguish normal patterns from abnormal ones. In contrast to most existing anomaly detection methods that only produce continuous anomaly score maps, we further design an adaptive threshold selection strategy for anomaly binarization. Specifically, building on the conventional statistical binarization rule of $\mu + k\sigma$, we cast threshold selection as the optimization of the coefficient k, which is tuned using synthetic anomalies for TEST_{priv} and further adapted with both lighting simulation and synthetic anomalies for $\mathrm{TEST}_{priv,mix}$, thereby yielding a more reliable threshold for binary mask generation. The resulting coarse masks are then further refined by HQ-SAM to improve boundary quality and segmentation accuracy. Experimental results on the MVTec AD 2 dataset demonstrate that our method achieves competitive performance across multiple object categories. Code is available at `isvl++`.

1. Introduction

Visual anomaly detection is a crucial component of quality assurance in modern manufacturing systems. By identifying deviations from the normal appearance of products, it improves both product reliability and process quality. [14]

Unsupervised anomaly detection (UAD) aims to identify abnormal patterns and localize anomalous regions using only normal training samples [6, 25]. This setting is particularly relevant to industrial inspection, where defective samples are often rare, diverse, or unavailable during training. By modeling the distribution of normal data, UAD methods can detect deviations without requiring exhaustive annotations of all possible defect types, making them well suited to real-world manufacturing scenarios.

Existing industrial anomaly detection methods can be broadly categorized into memory bank-based approaches [9, 10, 22], reconstruction-based approaches [1, 2, 12, 13, 16, 21] and distillation-based approaches [3, 5, 11, 20, 23]. Despite their architectural differences, most of these methods follow the same core principle: since only anomaly-free samples are observed during training, anomalous regions are expected to produce unusual feature responses or reconstruction errors that deviate from the learned normal distribution.

Although recent methods have achieved strong performance on standard benchmarks such as MVTec AD [4] and VisA [26], their deployment in real industrial environments remains challenging. Existing benchmarks often fail to capture the full complexity of practical inspection settings, including variations in illumination, background clutter, object appearance, and defect morphology. As a result, strong benchmark performance does not always translate into robust anomaly detection under real-world conditions.

To address these limitations, the recently introduced MVTec AD 2 dataset [15] provides a more challenging and realistic benchmark for industrial anomaly detection. It contains eight industrial scenarios with more than 8,000

*Main representative. Email: xingaowang@whu.edu.cn high-resolution images acquired under diverse lighting conditions. In comparison to earlier anomaly detection benchmarks, MVTec AD 2 presents a set of more challenging and practically relevant scenarios, such characteristics make the dataset a considerably more challenging testbed for assessing the robustness and real-world applicability of anomaly detection methods.

1.1. Challenge Description

We identify two main challenges when performing unsupervised anomaly detection on MVTec AD 2 [15].

Challenge 1: Robustness under varying industrial scenarios. MVTec AD 2 poses a substantially greater challenge than earlier anomaly detection benchmarks due to its wide range of industrial scenarios and pronounced variations in appearance, imaging conditions, and defect characteristics. In particular, the dataset includes transparent and reflective objects, overlapping structures, dark-field and backlight illumination, large intra-class variation among normal samples, and extremely small defects. These factors make it considerably more difficult to learn a stable model of normality from anomaly-free training data alone. Moreover, each scenario is evaluated under multiple lighting conditions, requiring anomaly detection methods to generalize under realistic distribution shifts rather than overfit to a specific acquisition setting.

Challenge 2: Adaptive anomaly binarization. Beyond robust anomaly scoring, practical industrial inspection also requires accurate binary defect masks. However, most existing anomaly detection methods produce continuous anomaly score maps instead of directly yielding binary predictions. This issue becomes even more pronounced on MVTec AD 2, where the score distribution can vary significantly across object categories and lighting conditions. Consequently, a fixed binarization rule often fails to produce reliable segmentation masks across different scenarios. This makes adaptive threshold selection a critical component for converting anomaly score maps into accurate binary predictions under varying imaging conditions.

2. Methodology

The pipeline of our method is illustrated in Fig. 1, which consists of three important parts: dataset pre-processing, model design and results post-processing.

2.1. Dataset pre-processing

Adaptive Overlap Cropping. MVTec AD 2 contains high-resolution product images, making direct training and inference computationally expensive. Meanwhile, defects often occupy only small local regions. To improve local anomaly sensitivity while reducing the computational burden, we adopt a dataset pre-processing strategy termed **Adaptive Sliding Crop**.

Specifically, each image is partitioned into fixed-size patches of 1024×1024 using an overlap-aware sliding window. Given a target overlap ratio ρ, which is 0.2 in the work, the nominal stride is defined as $s(1-\rho)$, where s denotes the window size. The crop positions along both horizontal and vertical directions are then uniformly adjusted such that the first and last windows align exactly with the image boundaries, thereby ensuring complete image coverage without missing peripheral regions. For images smaller than the predefined window size, boundary-replication padding is applied to maintain a consistent input size. The same cropping procedure is applied to the training, validation, and test sets. In addition, to preserve global contextual information, we also retain the original full-resolution training images and include them together with the cropped patches in the final training set.

Anomaly synthesis. We observe that structural anomalies in industrial inspection often exhibit recurring and visually similar patterns. In particular, two relatively stable defect types are frequently encountered: *linear defects*, such as cracks and scratches, and *blocky defects*, such as contamination and local material corruption. Based on this observation, we construct an anomaly library for synthetic data generation. Specifically, for *blocky defects*, which are relatively easy to simulate and exhibit diverse shapes, we randomly crop irregular patches from arbitrary images in external datasets, such as MVTec AD [4] and VisA [26], to serve as defect patterns. In contrast, *linear defects* are more challenging to simulate and tend to have relatively fixed shapes; for example, we crop anomalous regions from the crack category in the tile subset of MVTec AD and from several images in the screw category construct linear defect patterns.

To generate realistic pseudo-anomalous samples, we first extract the foreground region of each training image. This step is performed using SAM3 [7] with the corresponding category name as the text prompt. Since SAM3 does not always produce satisfactory foreground masks for all categories, we further employ the Foreground Estimating Branch proposed in CPR [18] as a fallback strategy for foreground extraction when necessary. We then paste anomaly patches from the anomaly library onto the extracted foreground regions to synthesize anomalous samples. The synthesis process is performed on the original-resolution training images to better preserve realistic defect scale and appearance, after which the synthesized images are resized to 640×640. Following this procedure, we generate 1024 synthetic anomalous images for each category, which are subsequently used for model training.

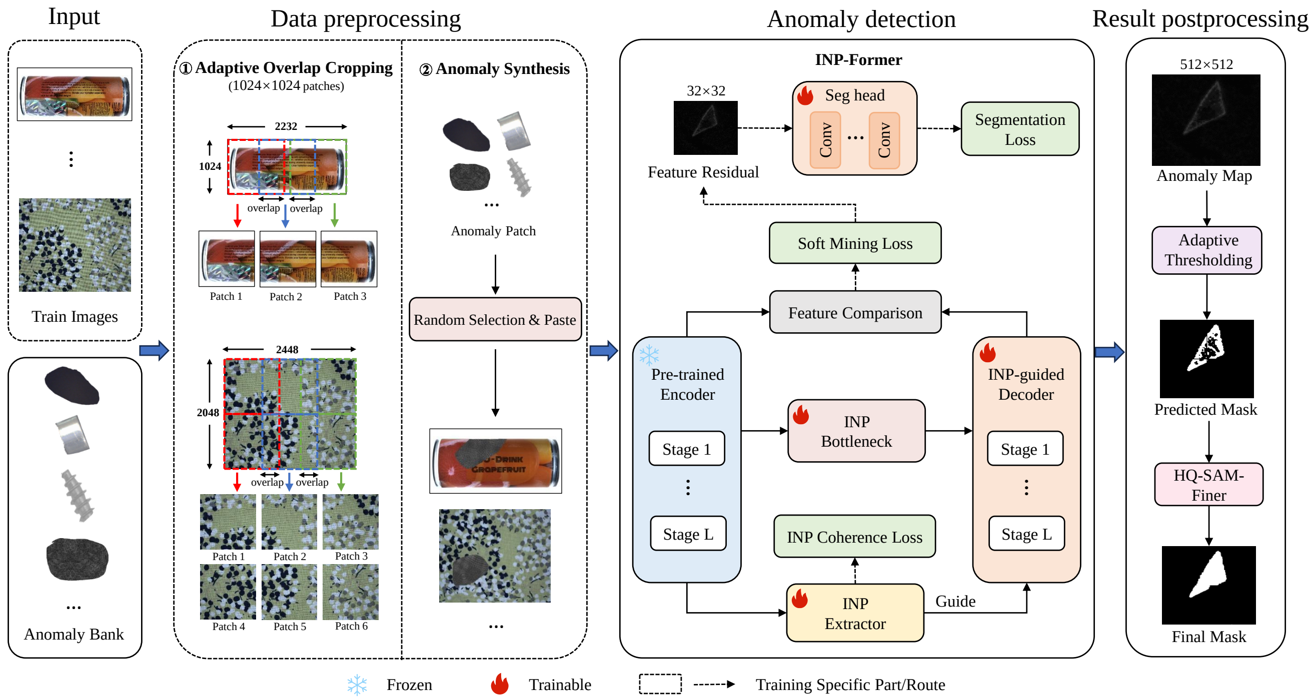

Figure 1. Our proposed framework, containing data preprocessing, anomaly detection and result postprocessing..

2.2. Model Design

2.2.1. Approach

We adopt INP-Former [21] as the backbone of our framework, motivated by the characteristics of MVTec AD 2. While some categories exhibit relatively regular local patterns, many others contain substantial intra-category sub-image variations, which requires the model to capture diverse normal visual structures in a flexible manner.

INP-Former is well suited to this setting, as its prototype-based design enables effective modeling of normal patterns and better generalization across heterogeneous sub-image variations. In our training pipeline, only normal training images are used as the data source. To further enhance the model's discriminative ability, we additionally construct synthetic anomalous samples from the normal images and use them as auxiliary supervision during training. In this way, the model learns to better distinguish normal patterns from abnormal perturbations without relying on real anomalous training samples. Moreover, we train one model for each product category, allowing the framework to better fit its category-specific normal distribution.

2.2.2. Architecture

Our framework is built upon INP-Former, with several modifications tailored to the challenge setting. The model mainly consists of a pretrained visual encoder, a prototype-guided reconstruction branch, and an additional residual-based segmentation branch.

Backbone encoder. We replace the original feature extractor with DINOv3 ViT-H/16 Plus [24], whose stronger representation capacity provides more discriminative features for industrial anomaly detection. To reduce computational cost while preserving sufficient defect details, each cropped sub-image of size 1024×1024 is resized to 512×512 before being fed into the network.

Prototype-guided reconstruction branch. For normality modeling, we retain the prototype-guided reconstruction paradigm of INP-Former and adapt it to the DINOv3 backbone. Specifically, we extract intermediate features from transformer blocks $\{8, 11, 14, 17, 20, 23, 26, 29\}$. To obtain more stable multi-level representations, we group adjacent layers into four feature groups, i.e., $\{8, 11\}$, $\{14, 17\}$, $\{20, 23\}$, and $\{26, 29\}$, for both encoder-side and decoder-side fusion. The reconstruction discrepancy between encoder and decoder features is then used as the primary anomaly cue.

Residual segmentation branch. Beyond the original INP-Former reconstruction pipeline, we introduce an additional residual segmentation head to improve anomaly localization. This branch operates on the multi-level discrepancies between encoder and decoder features. Specifically, for each fused feature level, we compute the absolute feature difference and the cosine dissimilarity, and concatenate them as complementary residual cues. These residual features are first projected into a shared hidden space and then concatenated for joint fusion.

To recover dense pixel-level predictions, we further employ a 4-stage upsampling decoder. Concretely, the fused residual representation is progressively upsampled through four transposed-convolution blocks, each followed by a convolution, batch normalization, and ReLU activation. The channel dimension is gradually reduced during this process, and a final 1×1 convolution is applied to produce a one-channel anomaly logit map at the target output resolution. During inference, the final anomaly map is obtained by a weighted combination of the reconstruction-based anomaly response and the probability map predicted by the residual segmentation head.

2.3. Training

We adopt a decoupled training strategy for the reconstruction branch and the residual segmentation branch. Following the challenge protocol, only normal images from the official training split are used as real training data, while synthesized anomalous images are introduced as auxiliary supervision for training only the residual segmentation head.

To improve robustness to diverse illumination conditions, we apply a photometric augmentation pipeline during training, including random exposure and gamma adjustment, directional illumination gradients, shadow and highlight simulation, vignetting, and mild additive noise. This augmentation encourages the model to learn illumination-invariant normal representations rather than overfitting to specific imaging conditions.

For the reconstruction branch, we optimize the trainable INP-Former modules using StableAdamW with a learning rate of 5×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 10^{-5}$, and $\epsilon = 10^{-10}$. The reconstruction objective combines the reconstruction loss and the gather loss with weights 1.0 and 0.2, respectively. We use a warmup cosine decay schedule, where the learning rate is warmed up for the first 100 iterations and then decayed from 5×10^{-4} to 1×10^{-5} over the full training process.

For the residual segmentation branch, we use synthesized anomaly images with binary masks for supervision. These samples are first forwarded through the current INP-Former to obtain encoder–decoder features, from which multi-level residual maps are constructed and fed into the segmentation head. Importantly, the segmentation loss is used only to optimize the residual head, while the reconstruction branch remains frozen by default. We optimize this branch with a separate StableAdamW optimizer using an initial learning rate of 3×10^{-4} and a final learning rate of 1×10^{-5}, while keeping the same β values, weight decay, and ϵ. The supervision is applied with a Dice and BCE loss. The segmentation branch starts to be optimized from epoch 1 and is updated once every two training iterations, with a warmup cosine decay schedule using 20 warmup iterations.

By default, we use a batch size of 8 for both normal training images and synthesized anomaly images, train the model for 12 epochs, and use up to 512 synthesized anomaly samples per category. In addition, gradient clipping with a maximum norm of 0.1 is applied to both optimization branches for stable training. During inference, the segmentation branch is assigned a fusion weight of 0.03 when combining its prediction with the reconstruction-based anomaly map.

2.4. Result Post-processing

During inference, our method predicts an anomaly map for each sub-image, where each pixel indicates the likelihood of being anomalous. To obtain the final binary segmentation mask for the original high-resolution image, we employ a three-stage post-processing pipeline.

(1) Anomaly map merging. Since inference is performed on sub-images, we first restore all anomaly predictions to the original image space. Specifically, each sub-image anomaly map is placed back to its corresponding position in the high-resolution image according to the cropping coordinates recorded during preprocessing. For overlapping regions between adjacent sub-images, the final anomaly score is computed by pixel-wise averaging. This procedure produces a merged anomaly map at the original image resolution.

(2) Category-adaptive thresholding. After merging, we convert the full-resolution anomaly map into an initial binary mask by thresholding. Following the strategy of the original MVTec AD 2 paper, we define the threshold for category c as

$$T_c = \mu_c + k_c \sigma_c,$$

where μ_c and σ_c denote the mean and standard deviation of the pixel-wise anomaly scores, respectively, computed on the merged full-resolution validation images. By default, we set $k_c = 5$.

However, since the optimal threshold coefficient may vary across categories, we further refine k_c using a synthetic validation protocol. Specifically, we construct pseudo-anomalous validation samples by anomaly synthesis with the DTD dataset [8], and then search for the category-specific optimal coefficient k_c^* on the synthesized validation set. The final coefficient is determined by averaging the default and data-driven choices:

$$k_c = \frac{5 + k_c^*}{2}.$$

The initial binary mask is then obtained by thresholding the merged anomaly map with T_c.

It is worth noting that TEST_{priv} and $\mathrm{TEST}_{priv,mixed}$ differ in lighting conditions, we perform an additional round of category-adaptive thresholding for $\mathrm{TEST}_{priv,mixed}$. Specifically, we incorporate lighting simulation during threshold calibration and re-optimize the category-specific threshold, which is then used to obtain the final results.

(3) HQ-SAM refinement. To further improve mask quality, following Robis [19], we adopt a SAM-Finer-style refinement strategy based on HQ-SAM [17]. Given the initial binary mask, we first identify its connected components and select at most the top three components according to region area. For each selected component, we extract its minimum bounding rectangle and use it as a box prompt for HQ-SAM to refine the corresponding abnormal region. The refined masks produced for all selected components are finally combined through a union operation to generate the final segmentation result. This refinement step improves boundary quality and helps recover more accurate abnormal regions from the coarse thresholded mask.

2.5. Dataset & Evaluation

We conduct experiments on MVTec AD 2. The complete dataset pre-processing workflow and technical details of our data augmentation strategies are documented in Section 2.1.

2.5.1. Evaluation Criteria

To evaluate the performance of our anomaly detection model, we used the official evaluation protocol. The primary metric is the F_1 score, which is the harmonic mean of precision and recall at the pixel level across different testing environments and objects in the MVTec AD 2 dataset.

The F_1 score is calculated as:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

where Precision is the ratio of true positive (TP) pixels to the total predicted positive pixels, and Recall is the ratio of TP pixels to the total actual positive pixels.

2.6. Evaluation Setting

We strictly adhere to the official regular setting of the challenge. Specifically, we do not use the MVTec AD 2 validation set for model training, nor do we use the test_public split for threshold selection.

3. Results

3.1. Performance Metrics

As shown in Tab 1, Our method achieves **58.31%** Seg-F1 on TEST_{priv} and **53.75%** Seg-F1 on $\mathrm{TEST}_{priv,mixed}$. Furthermore, it indicate that our approach performs robustly across various objects, with consistent performance even under changes in lighting. Notably, wallnuts achieved the highest score of 82.66% for TEST_{priv}. Objects such as fabric, fruit jelly, rice, vial, and walnuts show high $\mathrm{AucPRO}_{0.05}$ scores, consistently above 65%.

Table 1. Quantitative results on TEST_{priv} / $\mathrm{TEST}_{priv,mix}$ set.

Object	AucPro0.05	ClassF1	SegF1
Private			
Can	38.89	42.91	21.15
Fabric	74.24	83.16	88.55
Fruit Jelly	69.77	88.77	61.37
Rice	78.59	81.01	66.17
Sheet Metal	79.49	89.07	71.9
Vial	81.05	88.78	51.61
Wall Plugs	71.24	76.41	35.97
Walnuts	82.66	88.35	69.77
Mean	71.99	79.81	58.31
Private Mixed			
Can	31.52	56.45	3.55
Fabric	75.72	82.19	84.94
Fruit Jelly	69.59	89.51	61.32
Rice	78.89	80	66.84
Sheet Metal	79.98	87.24	69.47
Vial	77.22	86.65	49.96
Wall Plugs	57.89	74.09	24.6
Walnuts	78.61	81.72	69.29
Mean	68.68	79.73	53.75

Overall, the TEST_{priv} conditions yield slightly higher $\mathrm{AucPRO}_{0.05}$ scores than the $\mathrm{TEST}_{priv,mix}$ conditions, suggesting that lighting variations have a noticeable but manageable impact on the model's performance.

3.2. Comparison

As shown in Tab. 2, our method outperforms most baseline approaches across the majority of object categories on both official private test sets. Notably, compared with ISVL, the first-place submission of the previous edition, our unified framework achieves higher overall Seg-F1 by **4.50%** points on TEST_{priv} and **2.32%** points on $\mathrm{TEST}_{priv,mixed}$, despite not relying on multi-model ensembling. :contentReference[oaicite:1]index=1 This result highlights the effectiveness of our design and suggests that a single, well-structured framework can already outperform stronger ensemble-based competitors. In particular, our method achieves **35.97%/24.60%** Seg-F1 on Wall plugs, further demonstrating its advantage on challenging categories.

Table 2. Comparison of segmentation F_1 score (in %) across different methods on binarized images for TEST_{priv} / $\mathrm{TEST}_{priv,mix}$ set.

Object	Ours	ISVL	RD	RD++	EfficientAD	MSFlow	SimpleNet	DSR
Can	**21.15 / 3.55**	17.03 / 9.21	0.1 / 0.1	0.1 / 0.1	0.8 / 0.1	5.0 / 0.1	0.6 / 0.1	0.4 / 0.1
Fabric	**88.55 / 84.94**	84.95 / 81.76	2.6 / 2.2	2.9 / 2.3	7.6 / 1.0	22.0 / 4.1	21.6 / 10.2	7.9 / 5.0
Fruit Jelly	**61.37 / 61.32**	68.1 / 67.76	22.5 / 22.7	26.9 / 26.7	20.8 / 18.2	47.6 / 38.1	25.1 / 23.0	17.9 / 17.2
Rice	**66.17 / 66.84**	66.34 / 66.93	7.0 / 3.9	9.5 / 2.9	15.0 / 0.5	19.1 / 1.8	11.6 / 1.0	1.5 / 1.4
Sheet Metal	**71.9 / 69.47**	69.69 / 69.74	41.3 / 39.2	40.9 / 37.7	9.3 / 3.8	13.0 / 7.6	14.6 / 2.8	13.9 / 14.4
Vial	**51.61 / 49.96**	44.01 / 42.12	28.0 / 28.3	28.2 / 22.8	30.5 / 26.5	23.3 / 6.2	31.9 / 17.5	28.2 / 27.9
Wall Plugs	**35.97 / 24.6**	11.99 / 6.24	1.9 / 0.8	1.3 / 0.9	4.4 / 0.3	0.1 / 0.2	1.0 / 0.3	0.4 / 0.4
Walnuts	**69.77 / 69.29**	68.34 / 67.65	41.2 / 36.7	44.1 / 40.5	34.6 / 13.3	44.5 / 14.3	35.2 / 14.3	17.0 / 9.6
Mean	**58.31 / 53.75**	53.81 / 51.43	18.1 / 16.7	19.2 / 16.7	15.4 / 8.0	21.8 / 9.0	17.7 / 8.7	10.9 / 9.5

4. Discussion

4.1. Challenges & Solutions

Challenge 1: Robustness under varying industrial scenarios. A major challenge of the task is the large variation in appearance across industrial scenarios, including changes in illumination, object structure, and background complexity. To address this issue, we adopt a stronger visual backbone and further combine reconstruction-based anomaly detection with anomaly synthesis. The stronger encoder improves the representation of complex normal patterns, while the synthesized anomalies provide additional abnormal cues during training. Together, these designs enhance the model's robustness and improve its ability to detect anomalies under challenging real-world conditions.

Challenge 2: Adaptive anomaly binarization. Another key challenge is anomaly binarization. Since most methods produce continuous anomaly score maps, a fixed threshold often fails under different categories and lighting conditions. To address this issue, we adopt a statistical binarization rule of $\mu + k \cdot \sigma$ and cast threshold selection as the optimization of the coefficient k. Specifically, we tune k on synthesized anomalies from the validation set using DTD dataset, and further incorporate lighting simulation for $\mathrm{TEST}_{priv,mixed}$ to improve robustness under illumination shifts. This strategy yields more adaptive and reliable binary predictions across different test scenarios.

4.2. Model Efficiency

In terms of computational efficiency, our method achieves an average inference time of `1,064 ms` per image on an `RTX 4090` GPU. The runtime is estimated from the inference speed reported by the program progress bar during model execution, with the additional runtime introduced by the HQ-SAM refinement stage also taken into account.

4.3. Future Work

Although our method achieves promising performance overall, there remains substantial room for improvement on several challenging categories. We identify two main directions for future research.

First, improving the detection of small defects remains an important open problem. Based on our current results, *can* is the most challenging category, and its performance differs markedly between TEST_{priv} and $\mathrm{TEST}_{priv,mixed}$. A closer examination shows that the model tends to misclassify specular highlights as anomalies, especially under the more complex mixed-lighting setting. This observation suggests that future work should place greater emphasis on distinguishing true small-scale defects from illumination-induced artifacts, such as highlights and local reflections.

Second, our method could be improved by better balancing global context and local detail. In categories such as *fruit_jelly* and *vial*, the image typically contains only a single dominant object rather than repeated local patterns. In such cases, the intrinsic prototypes learned from cropped patches may be insufficient to fully characterize normality, while the full-image structure may provide more informative global cues. A promising direction is therefore to combine our current framework with memory-based methods, which represent normality using a memory bank of patch features and nearest-neighbor matching. Such a hybrid design may better integrate global semantic consistency with local anomaly sensitivity, and thus further improve performance on categories with limited structural repetition.

5. Conclusion

In this work, we presented a unified anomaly detection framework for the CVPR Workshop 2026 Visual Anomaly and Novelty Detection Challenge. By combining reconstruction-based anomaly detection with anomaly synthesis, our method achieves competitive performance on MVTec AD 2, reaching **58.31%** Seg-F1 on TEST_{priv} and **53.75%** Seg-F1 on $\mathrm{TEST}_{priv,mixed}$, and showing strong pixel-level segmentation ability in challenging industrial scenarios.

A key finding of this work is that synthetic anomalies are beneficial not only for improving the model's discriminative ability, but also for adaptive threshold selection. By casting statistical binarization as the optimization of the coefficient k in $\mu + k\sigma$, we showed that anomaly synthesis, together with lighting simulation, provides an effective way to obtain more reliable binary predictions under varying illumination conditions.

References

[1] Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. *arXiv preprint arXiv:1901.08954*, 2019. 1

[2] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018. 1

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 1

[4] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021. 1, 2

[5] Yunkang Cao, Qian Wan, Weiming Shen, and Liang Gao. Informative knowledge distillation for image anomaly segmentation. *Knowledge-Based Systems*, 248:108846, 2022. 1

[6] Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *arXiv preprint arXiv:2401.16402*, 2024. 1

[7] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts. 2026. 2

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 4

[9] Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1319–1329. IEEE, 2025. 1

[10] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pages 475–489. Springer, 2021. 1

[11] 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. 1

[12] Hewei Guo, Liping Ren, Jingjing Fu, Yuwang Wang, Zhizheng Zhang, Cuiling Lan, Haoqian Wang, and Xinwen Hou. Template-guided hierarchical feature restoration for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6447–6458, 2023. 1

[13] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20405–20415, 2025. 1

[14] Lars Heckler-Kram, Ashwin Vaidya, Jan-Hendrik Neudeck, Ulla Scheler, Dick Ameln, Samet Akcay, and Paula Ramos. From benchmarks to reality: Advancing visual anomaly detection by the vand 3.0 challenge. *arXiv preprint arXiv:2509.17615*, 2025. 1

[15] Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. *International Journal of Computer Vision*, 134(4):175, 2026. 1, 2

[16] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8526–8534, 2024. 1

[17] 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: 29914–29934, 2023. 5

[18] Hanxi Li, Jianfei Hu, Bo Li, Hao Chen, Yongbin Zheng, and Chunhua Shen. Target before shooting: Accurate anomaly detection and localization under one millisecond via cascade patch retrieval. *IEEE Transactions on Image Processing*, 33: 5606–5621, 2024. 2

[19] Xurui Li, Zhonesheng Jiang, Tingxuan Ai, and Yu Zhou. Robis: Robust binary segmentation for high-resolution industrial images. 2025. 5

[20] Yufei Liang, Jiangning Zhang, Shiwei Zhao, Runze Wu, Yong Liu, and Shuwen Pan. Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Processing*, 32:4327–4340, 2023. 1

[21] Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu.

Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9974–9983, 2025. 1, 3

[22] 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. 1

[23] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021. 1

[24] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 3

[25] Guoyang Xie, Jinbao Wang, Jiaqi Liu, Jiayi Lyu, Yong Liu, Chengjie Wang, Feng Zheng, and Yaochu Jin. Im-iad: Industrial image anomaly detection benchmark in manufacturing. *IEEE Transactions on Cybernetics*, 54(5):2720–2733, 2024. 1

[26] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In *European conference on computer vision*, pages 392–408. Springer, 2022. 1, 2