AURD: Final submission in VAND3

Xu Tan[1] Chaohui Chen[1] Haidong Gao[2] Deyang Nan[1] Yang Yu[1]

[1](School of Computer Science and Technology, Zhejiang University of Science and Technology
[2](School of Computer Science and Technology, Zhejiang University of Technology

Contact Information: tanxu@zust.edu.cn
Track: Track I - Adapt & Detect

Abstract In this project, we introduce Augmented URD (AURD), an enhanced version of the URD[1] framework for anomaly detection based on the Expert-Teacher-Student (E-T-S) network architecture. AURD integrates an improved image augmentation pipeline, a refined training strategy, and a segmentation network that fuses multi-level feature differences in a trainable manner, enabling robust anomaly detection in real-world scenarios.

By applying various augmentations to generate synthetic anomalies, AURD simulates a wide range of potential anomalies that may occur in practical scenes. In AURD, foreground objects are precisely isolated from the background, enabling accurate addition of noise to these objects to synthesize anomalies.

Additionally, AURD incorporates a segmentation network that uses generated binary anomaly masks as supervisory signals during training to guide feature fusion. This approach replaces empirical aggregation methods (such as summation and weighted averaging) with a learned, trainable fusion strategy that adaptively adjusts weights based on the characteristics of different tasks and datasets.

Our empirical evaluations demonstrate that AURD achieves excellent performance across various categories, the pixel-level F1 scores of 45 and 41.2 were achieved on the private and private_mixed datasets, respectively.

1 Introduction

1.1 Background

In the industrial domain, Visual Anomaly Detection (VAD) plays a crucial role in identifying defects in the manufacturing process. Accurate anomaly detection not only ensures the quality and reliability of products but also reduces waste and recall costs, thereby optimizing production expenses. The rapid advancement of machine learning has revitalized this field, with supervised, unsupervised, and semi-supervised learning methods driving significant technological progress. Supervised learning algorithms, trained on large labeled datasets, have shown impressive performance in anomaly detection tasks. However, in real-world industrial settings, obtaining high-quality labeled data is often costly and time-consuming,

which limits the widespread adoption of supervised methods.

Against this backdrop, one-class classification models—which typically learn from normal data—have emerged as a prominent focus in unsupervised or semi-supervised anomaly detection research. These models build a representation of normal behavior by learning the feature distribution of non-defective samples, identifying deviations as anomalies. This approach effectively bypasses the challenge of insufficient labeled data. However, the complexity and variability of industrial production environments make data distributions prone to change due to numerous factors, resulting in domain shift. For instance, replacing image acquisition devices may introduce differences in camera parameters, and fluctuating lighting conditions can also alter the visual characteristics of data. These shifts can undermine the generalization capability of anomaly detection models, increasing false alarms and missed detections, and ultimately compromising quality control.

Therefore, investigating methods to improve the adaptability and robustness of visual anomaly detection models under domain shift has become a critical scientific challenge. Addressing this issue is essential not only for enhancing industrial production quality and efficiency, but also for supporting the development of intelligent and autonomous industrial automation.

1.2 Challenge Description

The Visual Anomaly and Novelty Detection (VAND) Challenge aims to address the limitations of existing systems in handling real-world conditions not represented in the training set. Its core objective is to develop robust models with strong generalization capabilities, enabling them to adapt to unpredictable domain changes. Through innovative algorithms and model optimization, participants are encouraged to design systems that can dynamically adapt to evolving data distributions and accurately identify novel anomalies and out-of-distribution samples.

The challenge leverages the new MVTec Anomaly Detection 2 (MVTec AD 2) [2] dataset, which includes eight challenging real-world scenarios captured under varying illumination conditions to reflect real-world distribution shifts. Participants are tasked with developing models based on the one-class training paradigm, where training is conducted solely on normal images. This approach is essential for ensuring that models can generalize to unseen anomalies without prior exposure to specific defect types.

The test set includes a collection of undisclosed perturbations that simulate real-world variations and noise, such as changes in camera angles, lighting conditions, and environmental interference. The challenge uses the pixel-level F1 score (SegF1) as its primary evaluation metric, balancing precision and recall in anomaly detection performance. Furthermore, participants are required to select a single threshold for binarizing the typically continuous anomaly maps —a task often overlooked in academic research but critical for real-world deployment.

2 Methodology

2.1 Approach

To address the challenge of robust anomaly detection, we began by selecting a suitable model through a comprehensive evaluation of state-of-the-art methods. By analyzing the strengths and limitations of various architectures, we identified an appropriate baseline for further development. Currently, three leading approaches dominate the field: patch-matching methods, reconstruction methods, and student-teacher frameworks. We examined PatchCore[3] as a representative of patch-matching approaches, SimpleNet[4] for reconstruction-based methods, and URD[1] for student-teacher models.Patch-based methods rely on a large memory bank to store diverse features, which can significantly slow down both training and inference. Deep learning-based reconstruction models often require extensive training time—ranging from hours to days—when applied to high-resolution images or large-scale datasets, and they demand considerable computational resources during inference. Based on these observations, we chose the student-teacher paradigm for its balance between efficiency and accuracy, using URD[1] as the foundation.We extended URD by implementing enhanced training strategies and incorporating a segmentation network module. As a result, we named our improved framework Augmented-URD (AURD).

2.2 Architecture

We propose a method based on URD[1], which leverages the power of synthetic anomaly images during training. The architecture consists of an Expert-Teacher-Student (E-T-S) network and a bottleneck module. The bottleneck includes a Multi-scale Feature Fusion (MFF) module and a One-Class Embedding (OCE) module. A Guidance Information Injection (GII) module is inserted before these two modules within the student network.The expert encoder E and the teacher encoder T are WideResNet50 models pre-trained on ImageNet. The expert network's parameters are frozen, while the teacher, student, and bottleneck networks remain trainable. Knowledge is distilled from the expert network to both the teacher and student networks simultaneously. This distillation enhances the teacher' s sensitivity to anomalies and reduces noise in the student' s feature representations. As a result, the teacher and student networks exhibit similar features in normal regions but diverge in anomalous regions, enabling effective anomaly detection and localization.The Guidance Information Injection (GII) module employs a similarity-based attention mechanism derived from higher layers to guide the transfer of encoder information to the decoder. This not only compensates for missing low-level information during reconstruction but also suppresses residual anomaly signals in the decoding process, thereby preventing anomaly leakage.

We also incorporate a segmentation network[5], as illustrated in Figure 1. Traditionally, anomaly detection methods sum cosine distances across multiple feature levels to compute pixel-wise anomaly scores. However, when different feature levels exhibit varying discriminative power, simple summation may yield suboptimal results. To address this, we guide feature fusion using binary anomaly masks generated from synthetic anomalies.The segmentation network consists of two residual blocks followed by an Atrous Spatial Pyramid Pooling (ASPP) module. During training, the weights of the teacher and student networks are

frozen to prevent interference, and only the segmentation network is optimized. Synthetic anomaly images are input to the teacher-student (T-S) network, with corresponding binary masks used as ground truth for supervision.We train the segmentation module using a combination of focal loss and L1 loss. Focal loss helps the model focus on minority and difficult samples, while L1 loss sharpens the boundaries of the predicted segmentation masks, improving localization accuracy.

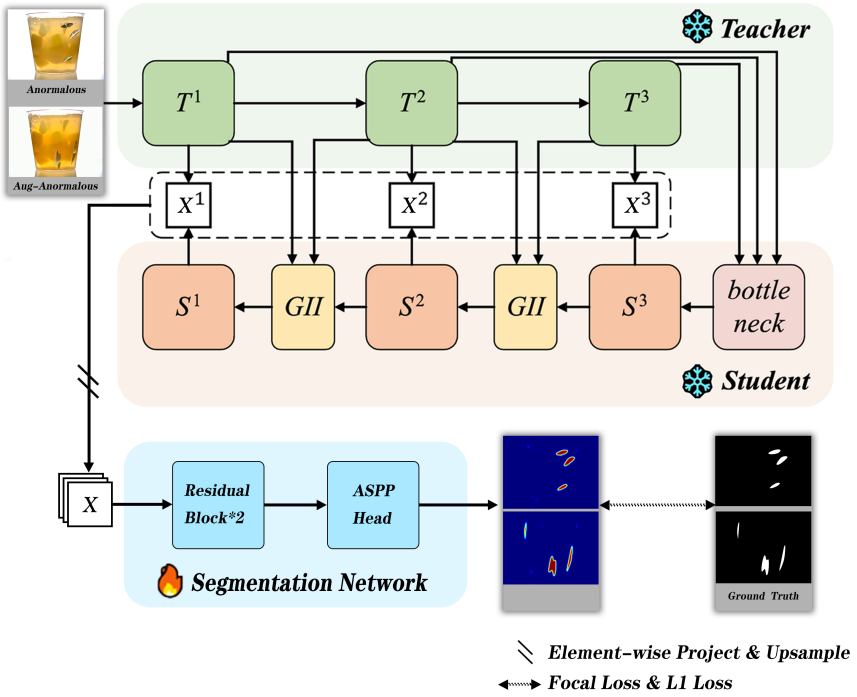

Figure 1 Segmentation Network Module Diagram

The primary objective of image augmentation is to help the model adapt to potential real-world domain shifts during inference. During AURD training, we apply a variety of augmentations, including brightness, contrast, and saturation adjustments to simulate different lighting conditions and times of day that may vary across deployment environments. Gaussian noise and blur are introduced to mimic sensor noise and defocus effects, respectively. For texture categories, two additional geometric transformations—flipping and rotation—are applied to simulate variations in camera angles. Flipping, in particular, further enhances image diversity. However, in object categories with fixed object placement, applying flipping and rotation may result in unintended object displacement or disappearance.Additionally, we incorporate two-dimensional Perlin noise to synthesize more realistic anomalies. Figure 2 illustrates examples of the augmented images produced through these transformations.

2.3 Training

We employ a one-class training paradigm, training a separate model for each class. Each model is trained independently on a single RTX 5090 GPU. To ensure compatibility with the network architecture, we resize input images such that the shorter side is reduced to one-fourth of its original length, and the

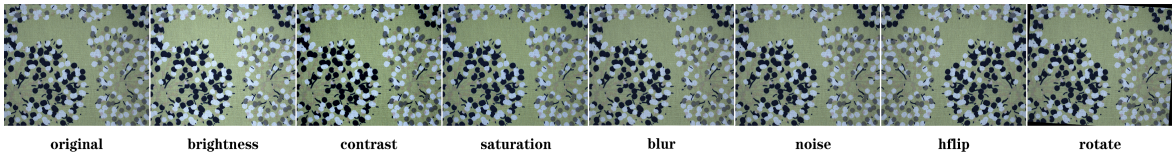

Figure 2 The examples of augmented images used

longer side is scaled proportionally. Afterward, the image dimensions are adjusted to ensure both height and width are multiples of 32. The training batch size is set to 4. We apply an early stopping strategy, with a hard cap of 10,000 training iterations. The bottleneck layer and the student network share an Adam optimizer with a learning rate of 0.005. The teacher encoder is trained separately using another Adam optimizer with a learning rate of 0.0001. The segmentation network is optimized using SGD with a learning rate of 0.001.

3 Dataset Evaluation

The MVTec Anomaly Detection 2 (MVTec AD 2) dataset[2] comprises eight challenging real-world scenarios, each corresponding to a distinct object category. The training set contains only normal images, while the test set includes various types of defects. Additionally, the test set introduces environmental lighting variations to simulate real-world conditions and evaluate model robustness.

We evaluated our model's performance using the pixel-level F1 score, which provides a balanced measure of precision and recall for comprehensive anomaly detection assessment.

4 Results

Table 1 presents the pixel-level F1 metrics on the test datasets test_private and test_private_mixed, comparing our model with URD[1]. We observe that the metrics of our model have improved, indicating the effectiveness of our approach. Among them, aug_URD refers to the use of augmented images for training.

Table 1 Pixel-Level F1 Score Results on the test_private and test_private_mixed Datasets

Data	Object	URD[1]	aug_URD	Ours	Data	URD[1]	aug_URD	Ours
test_private	Can	5.5	7.73	7.89	test_private_mixed	0.9	2.65	2.42
	Fabric	9.45	48.23	49.53		7.44	39.94	50.43
	FruitJelly	51.98	41.87	64.32		50.25	42.04	62.91
	Rice	32.69	39.42	57.99		29.72	39.33	46.9
	SheetMetal	49.53	49.96	56.09		52.3	53.34	58.78
	Vial	38.71	39.3	42.56		38.42	38.41	43.78
	Wallplugs	17.24	18.76	22.08		7.05	7.28	8.3
	Walnuts	56.89	55.89	59.51		55.4	56.06	56.12
	Average	32.75	33.77	45		30.18	31.27	41.2

5 Discussion

When conducting visual anomaly detection in dynamic environments, our goal is to identify anomalous pixels within images. The primary challenge in this task is distribution shift. Most existing anomaly detection methods assume that the training and test data are drawn from the same distribution. However, in practical applications, factors such as changing lighting conditions or background variations often cause significant shifts in the test data distribution. This can lead to the failure of traditional anomaly detection methods, which rely on a consistent data distribution. To mitigate this, we have implemented various data augmentation techniques, helping the model generalize to previously unseen scenarios.

In future work, we plan to explore more efficient anomaly detection and data augmentation strategies to better handle distribution shifts.

6 References

[1] LIU X, WANG J, LENG B, et al. Unlocking the potential of reverse distillation for anomaly detection [C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 39. 2025: 5640-5648.

[2] HECKLER-KRAM L, NEUDECK J H, SCHELER U, et al. The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection[A]. 2025.

[3] ROTH K, PEMULA L, ZEPEDA J, et al. Towards total recall in industrial anomaly detection[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 14318-14328.

[4] LIU Z, ZHOU Y, XU Y, et al. Simplenet: A simple network for image anomaly detection and localization [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 20402-20411.

[5] ZHANG X, LI S, LI X, et al. Destseg: Segmentation guided denoising student-teacher for anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 3914-3923.